

Serial No. 09/020,716
Group Art Unit: 1638

110. (Amended) A transformed maize seed which has been transformed with a plant polynucleotide to express a polypeptide in the endosperm of the transformed maize seed, wherein the transformed maize seed exhibits an elevated level of lysine or a sulfur-containing amino acid compared to a corresponding non-transformed maize seed.

REMARKS

Reconsideration of the present application is respectfully requested.

Claims 76-79, 90-93 and 95-111 are pending in the application. As discussed in detail below, the claims have been amended to delete certain words objected to by the Examiner.

Claim 104 is rejected under 35 USC 112, first paragraph, as containing subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventors, at the time the application was filed, had possession of the claimed invention. The Examiner states that there does not appear to be support in the specification for the specific mole % recited in the claim.

The Examiner's attention is drawn to page 6, lines 14-21 of the present application. High lysine content protein and high sulfur content protein are described in the specific terms found in claim 104. However, in order to expedite prosecution claim 104 has been amended to delete "to about 50 mole %" and "to about 40 mole %". "At least" has been added before about 7 mole % and about 6 mole %. Support for the amendment is found in the same location in the application.

Claims 76-79, and 90-93 remain rejected and new claims 95-111 are rejected under 35 USC 112, first paragraph, as containing subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventors, at the time the application was filed, had possession of the claimed invention.

Serial No. 09/020,716
Group Art Unit: 1638

The rejection is respectfully traversed. The arguments in the previous responses are maintained. The Examiner states that although the specification refers to other wild type polypeptides, Applicant does not describe other modified nucleic acids nor plants comprising said nucleic acids that have increased lysine or sulfur-containing amino acids. The Examiner invites Applicants to submit copies of references published prior to the filing date of the present application that teach other nucleic acid molecules that could be used in the claimed method to increase lysine or sulfur containing amino acids in plants.

As discussed in detail below, numerous wild-type and modified polynucleotides are disclosed in the application and are also known in the art. Copies of publications in addition to those previously provided are submitted with this response.

The Examiner states that it is improper to incorporate essential material by reference and that the Applicant has not satisfied the written description requirement.

It is respectfully submitted that particular polynucleotide sequences are not critical to the broad claims. In fact it would be impossible to submit all possible sequences that could be used in the claims. Claim 78 calls for a polynucleotide that encodes HT12 or ESA. These sequences were filed with the original application as SEQ ID NOS: 2 and 6 respectively as discussed below.

As requested by the Examiner, copies of references discussed below will be provided unless they were already submitted in a 1449 form. The location of the polynucleotide sequences can readily be determined in the various publications. These references demonstrate the skill in the art with regard to polynucleotides that encode proteins with elevated levels of lysine or sulfur-containing amino acids. If additional publications are needed, they can be provided by Applicant.

With regard to the ESA nucleic acid, the sequence is found in SEQ ID NO: 6, (2199-2675) (see Table 2, page 40, of the present application). The hordothionin

Serial No. 09/020,716
Group Art Unit: 1638

(HT) SEQ ID NO: 1, (3361-2947), high lysine hordothionin (HT12) SEQ ID NO: 2 (3361-2947) and the high lysine chymotrypsin inhibitor gene (also called barley high lysine gene or BHL) SEQ ID NO. 7 (2199-2450) are found in the sequences filed and identified in Table 2 of the present application. Additional HT12 sequence modifications are found in SEQ ID NOS: 10-13.

In addition numerous suitable genes were known in the art, many identified in the application. The Examiner is familiar with the Rao patents as they were cited in 1449 forms. US Ser. No. 08/838,763 cited on page 8, line 23 of the present application is now US Pat. No. 5,990,389, cited on a 1449 form as A18. US Ser. No. 08/824,379 cited on page 8, line 24 of the present application is now US Pat. No. 5,885,801 cited on a 1449 form as A20. US Ser. No. 08/824,382 cited on page 8, line 24 of the present application is now US Pat. No. 5,885,802, cited on a 1449 form as E2. The 10 kD zein storage protein from maize is disclosed in Kiriwara et al. 1988, Mol. Gen. Genet. 211: 477-484, a copy of which is enclosed. Sulfur-rich 10 kD rice prolamin is disclosed in Masumura et al., Plant Mol. Biol. 12: 123-130, 1989, (A25 on the 1449 form and cited on page 13, lines 7-8 of the present application, SEQ ID NOS: 20-21). The maize gene encoding methionine-rich 15 kD zein protein is found in Pedersen et al., J. Biol. Chem., 261, 6279-6284 (1986), (A26 on the 1449 form and cited on page 13, lines 5-6 of the present application, SEQ ID NOS: 16-17). The gene encoding the Brazil nut protein is found in Altenbach et al., Plant Mol. Biol., 8: 239 (1987), a copy of which is included. The gene encoding a high methionine maize 10 kD zein is found in Kiriwara et al., Gene, 7, 359-370 (1988), (A22 on the 1449 form submitted and cited on page 13, lines 6-7 of the present application). Pea genes encoding high sulfur protein are disclosed in Higgins et al., J. Biol. Chem., Vol. 261, No. 24, pp. 11124-111310 (1986), (A21 on the 1449 form and cited on page 12, lines 6-7 of the present application, SEQ ID NOS: 14-15). A gene encoding a methionine rich sunflower protein is found in Lilley, et al., Proceedings of the World Congress on Vegetable Protein Utilization in Human

Serial No. 09/020,716
Group Art Unit: 1638

Foods and Animal Feedstuffs; Applewhite, T.H. (ed.), American Oil Chemists Soc., Champaign, IL, pp. 497-502 (1989), (A23 on the 1449 form and cited on page 13, lines 1-5 of the present application).

Other suitable genes include 12S seed storage protein gene from rapeseed disclosed in Ryan et al., *Nucleic Acids Res.*, 17 (9): 3584 (1989) a copy is enclosed. The sunflower 2S albumin gene is disclosed in Allen et al., *Mol. Gen. Genet.*, 201 (2): 211-218, (1987) a copy is enclosed. The maize albumin b-32 gene is disclosed in Di Fonzo et al., *Mol. Gen. Genet.*, 212 (3): 481-487 (1988), a copy is enclosed. The napin gene is disclosed in Joseffson et al., *J. Biol. Chem.*, 262 (25): 12196-12201 (1987) and Scofield and Couch, *J. Biol. Chem.*, 262 (25): 12202-12208 (1987) copies are enclosed. The B1 hordein gene is disclosed in Forde et al. *Nucleic Acids Res.* 13 (20): 7327-7339 (1985), a copy is enclosed. The wheat alpha and beta gliadin genes were described in Sumner-Smith et al., *Nucleic Acids Res.*, 13 (11): 3905-3916 (1985), a copy is enclosed. Wheat gliadin is also disclosed in Anderson et al., *Nucleic Acids Res.*, 12(21): 8129-8144 (1984), a copy is enclosed. The pea legumin gene is disclosed in Lycett et al., *Nucleic Acids Res.*, 12 (11): 4493-4506, a copy is enclosed. Various maize zeins are disclosed in Heidecker and Messing, *Nucleic Acids Res.*, 11 (14): 4891-906 (1983), copies are enclosed. The alpha, alpha', and beta-subunits of soybean 7S seed storage protein is disclosed in Schuler et al., *Nucleic Acids Res.*, 10 (24): 8245-8261 (1982) and Schuler et al., *Nucleic Acids Res.*, 10 (24) 8225-8244 (1982) copies are enclosed. The sunflower 11S gene is described in Vonder Haar et al., *Gene*, 74 (2): 433-443 (1988), a copy is enclosed. The pea convicilin gene is disclosed in Bown et al., *Biochem. J.*, 251 (3): 717-726 (1988), a copy is enclosed.

Claims 76-79, and 90-93 remain rejected and new claims 95-111 are rejected under 35 USC 112, first paragraph, because the specification is enabling only for claims limited to transformed cereal plant seed having an elevated lysine, methionine and cysteine content (about 10% to about 35%) by weight compared to

Serial No. 09/020,716
Group Art Unit: 1638

untransformed cereal plant seed) comprising the modified hordothionin gene of SEQ ID NO: 2 (HT12), vectors, plant cells and transformed plants comprising said modified hordothionin gene. The Examiner states that the specification does not enable any person skilled in the art to which it pertains, or with which it is most nearly connected to make and or use the invention commensurate in scope with these claims.

The rejection is respectfully traversed. As discussed above, numerous useful genes are cited in the application. Many others were known at the time of filing. Further a 1.132 Declaration was submitted October 18, 1999 by Rudolf Jung, a co-inventor on the application. The results in the Declaration demonstrate significant increases in the level of methionine when using ESA as the polynucleotide. Increases in the level of methionine of up to 30 % were demonstrated.

The Examiner states that claim 104 is not enabled for 50 mole % lysine or 40 mole % sulfur.

In order to simplify the claim and expedite prosecution, claim 104 has been amended to remove "50 mole % lysine" and "40 mole % sulfur".

Claims 76-79, 90-93, and 95-111 are rejected under 35 USC 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

Claims 76 and 77 have been amended as suggested by the Examiner to recite "transformed cereal plant" rather than "transformed cereal plant seed".

Claims 98 and 99 have not been amended in a similar fashion because there is no antecedent basis for "transformed cereal plant".

The Examiner objects to the phrase "plant derived polynucleotide" claims 73, 95-97, 104-108 as there are many types of derivatives and hence it is not known what is encompassed by derived.

Serial No. 09/020,716
Group Art Unit: 1638

The claims have been amended as suggested by the Examiner to remove "derived" from the claims. The amended claims read a "plant polynucleotide". The claims encompass plant polynucleotides as described throughout the specification.

Claims 101 and 102 are objected to because of the phrase "about 10 times" is considered indefinite. The phrase has been removed to expedite prosecution.

Claims 76-79 and 90-93 remain rejected and new claims 95-111 are rejected under 35 USC 102(e) as being anticipated by Falco et al. (U.S. Patent 5,773,691).

The Examiner states that in view of the indefinite claim language "plant derived polynucleotide", it reads on essentially any polynucleotide, because any polynucleotide can be "derived" from a plant. As noted above, the claims have been amended to remove "derived". The amended claims require a "plant polynucleotide".

The Examiner further states that Falco teaches plant polynucleotides in Example 20.

It is noted that the LKR gene of Example 20 is an enzyme that is involved in lysine catabolism. In order to increase lysine one needs to suppress expression of the LKR. If LKR is expressed the level of lysine is decreased. Therefore, Example 10 does not anticipate the present claims, which require expression of a polypeptide.

Claims 76-79 and 90-93 remain rejected and new claims 95-111 are rejected under 35 USC 103(a) as being unpatentable over Rao et al. (US Patent 5,885,802) in view of Applicant's admission and also over Rao et al. (US Patent 5,990,389).

The Examiner states that substitution of one promoter for another promoter is routine in the art.

The rejection is traversed and the previous arguments are maintained. Namely, there is no motivation or suggestion in the art to use an endosperm preferred promoter or that it would produce beneficial results.

Serial No. 09/020,716
Group Art Unit: 1638

The Examiner states that the Falco teaching cannot be considered because Applicant has not cited a reference or a location in that reference for the quotation of Falco. The reference and location are cited below.

In US 5,773,691, Example 26, Col. 88, Lines 34-41, Falco et al. state "No increase in free lysine was observed in seed expressing *Corynebacterium* DHDPS plus *E. coli* from the glutelin 2 promoter with or without AKIII-M4". Falco et al. further indicate that "lysine catabolism is expected to be much greater in the endosperm than the embryo and this probably prevents the accumulation of increased levels of lysine in seeds expressing *Corynebacterium* DHDPS plus *E. coli* AKIII-M4 from the glutelin 2 promoter".

The DHDPS gene expressed by glutelin 2 (an endosperm preferred promoter) did not increase lysine in the seed. Falco et al. concluded that lysine catabolism is greater in the endosperm, thus preventing an increase in lysine. Falco et al. therefore teach away from the present claims. The present claims require an endosperm preferred promoter and/or expression of a polypeptide in endosperm. The Supreme Court held in *US v Adams*, 383 US 39, 148 USPQ 479 (1966) that one important indicia of nonobviousness is "teaching away from the claimed invention by the prior art or by experts in the art at (and/or after) the time the invention was made. The decision maker must consider the prior art as a whole in making an obviousness rejection. Also see *In re Fine*, 837 F.2d 1071, 5 USPQ2d 1596 (Fed. Cir. 1988). Teaching away from the art is a per se demonstration of lack of prima facie obviousness. There can be no expectation of success. The prior art as a whole must be considered. To proceed contrary to accepted wisdom is strong evidence of nonobviousness. *In re Hedges*, 228 USPQ 685, 687 (Fed. Cir. 1986).

In 35 USC 103, the statute expressly requires that obviousness or nonobviousness be determined for the claimed subject matter as a whole. The results and advantages produced by claimed subject matter must be considered. As

Serial No. 09/020,716
Group Art Unit: 1638

discussed above, the results and advantages were not disclosed or suggested in the prior art. *Diversitech Corp. v. Century Steps, Inc.* 7 USPQ2d 1315 (Fed. Cir. 1988).

The Examiner states that the motivation combining the elements of the present invention is provided in the Rao reference itself. The Examiner further states that Rao shows increases in amino acid composition in the seed (the major portion of which is the endosperm) with the constitutive promoter, one would have been motivated to substitute a seed-specific, or endosperm-specific promoter to further increase or to limit increases to the seed/endosperm tissue. The Examiner concludes that it would have been an obvious modification to substitute an endosperm-specific promoter.

It is again emphasized that there must be some motivation to make the particular claimed combination. There are many possible types of promoters to choose from. There was no motivation to choose endosperm preferred promoters.

Claims 76-79 and 90-93 remain rejected and new claims 95-111 are rejected under 35 USC 103(a) as being unpatentable over Jaynes et al. (US pat. 5,811,654) in view of Applicant's admission. The Examiner states that the teachings of Jaynes are clearly directed to increasing amino acid compositions in seed and that it would have been an obvious modification to substitute an endosperm-specific promoter.

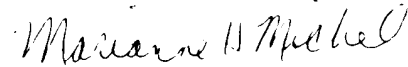
The rejection is respectfully traversed. Arguments in the previous responses are maintained. In particular, there is no suggestion or motivation to make the claimed combination. As discussed in detail above Falco teaches away from the using an endosperm-specific promoter. Based on the prior art at the time of filing, one would have no expectation of success when using an endosperm preferred promoter to increase the level of amino acid in a seed.

Attached hereto is a marked-up version of the changes made to the specification by the current amendment. The attached page is captioned **"Version with markings to show changes made."**

Serial No. 09/020,716
Group Art Unit: 1638

In view of the above comments and amendments, withdrawal of the outstanding rejections and allowance of the remaining claims is respectfully requested.

Respectfully submitted,



Marianne H. Michel
Attorney for Applicant
Registration No. 35,286

PIONEER HI-BRED INTERNATIONAL, INC.
Corporate Intellectual Property
7100 N.W. 62nd Avenue
P.O. Box 1000
Johnston, Iowa 50131-1000
Phone: (515) 334-4467
Facsimile: (515) 334-6883

VERSION WITH MARKINGS TO SHOW CHANGES MADE

In the claims:

76. (Twice amended) The method of claim 95, wherein the transformed cereal plant [seed is from] is maize, wheat, rice, or sorghum.
77. (Twice Amended) The method of claim 76 wherein the transformed cereal plant [seed is from] is maize or sorghum.
95. (Amended) A method for increasing the level of lysine or a sulfur-containing amino acid in a cereal plant seed, the method comprises transforming a cereal plant cell with an expression cassette and regenerating a transformed cereal plant to produce a transformed cereal plant seed, wherein the expression cassette comprises a seed endosperm-preferred promoter operably linked to a plant [derived] polynucleotide encoding a polypeptide, and wherein expression of the polypeptide increases the level of lysine or a sulfur-containing amino acid in the transformed cereal plant seed compared to a corresponding non-transformed cereal plant seed.
96. (Amended) The method of claim 95 wherein the seed endosperm-preferred promoter is heterologous to the plant [derived] polynucleotide.
97. (Amended) A transformed cereal plant seed which has been transformed with a plant [derived] polynucleotide to express a polypeptide in endosperm of the transformed cereal plant seed, wherein the transformed cereal plant seed exhibits an elevated level of lysine or a sulfur-containing amino acid compared to a corresponding non-transformed cereal plant seed.

101. (Amended) The transformed cereal plant seed according to claim 100 wherein the amount of lysine or sulfur-containing amino acid in the transformed cereal plant seed is increased at least about 15 percent by weight [to about 10 times] compared to a corresponding non-transformed cereal plant seed.
102. (Amended) The transformed cereal plant seed according to claim 101 wherein the amount of lysine or sulfur-containing amino acid in the transformed cereal plant seed is increased at least about 20 percent by weight [to about 10 times] compared to a corresponding non-transformed cereal plant seed.
104. (Amended) An expression cassette comprising a seed endosperm-preferred promoter operably linked to a plant [derived] polynucleotide encoding a polypeptide having at least about 7 mole % [to about 50 mole %] lysine or at least about 6 mole % [to about 40 mole %] of a sulfur containing amino acid.
105. (Amended) The expression cassette of claim 104 wherein the seed endosperm-preferred promoter is heterologous to the plant [derived] polynucleotide.
106. (Amended) A seed from a transformed cereal plant which has been transformed with a plant [derived] polynucleotide to express a polypeptide in the endosperm of the transformed cereal plant seed, wherein the transformed cereal plant seed exhibits an elevated level of lysine or a sulfur-containing amino acid compared to a corresponding non-transformed cereal plant seed.

107. (Amended) A method for increasing the level of lysine or a sulfur-containing amino acid in a maize seed, the method comprises transforming a maize cell with an expression cassette and regenerating a transformed maize plant to produce a transformed maize seed, wherein the expression cassette comprises a seed endosperm-preferred promoter operably linked to a plant [derived] polynucleotide encoding a polypeptide, and wherein expression of the polypeptide increases the level of lysine or a sulfur-containing amino acid in seed of the transformed maize plant compared to seed of a corresponding non-transformed maize plant.
108. (Amended) The method of claim 107 wherein the seed endosperm-preferred promoter is heterologous to the plant [derived] polynucleotide.
110. (Amended) A transformed maize seed which has been transformed with a plant [derived] polynucleotide to express a polypeptide in the endosperm of the transformed maize seed, wherein the transformed maize seed exhibits an elevated level of lysine or a sulfur-containing amino acid compared to a corresponding non-transformed maize seed.

attachment
#37

Differential expression of a gene for a methionine-rich storage protein in maize

Julie Anderson Kiriwara¹, John P. Hunsperger², Walter C. Mahoney^{2*}, and Joachim W. Messing¹

¹Waksman Institute, Rutgers, The State University, Piscataway, NJ 08855, USA

²Department of Genetics and Cell Biology, University of Minnesota, St. Paul, MN 55108, USA

Summary. A methionine-rich 10 kDa zein storage protein from maize was isolated and the sequence of the N-terminal 30 amino acids was determined. Based on the amino acid sequence, two mixed oligonucleotides were synthesized and used to probe a maize endosperm cDNA library. A full-length cDNA clone encoding the 10 kDa zein was isolated by this procedure. The nucleotide sequence of the cDNA clone predicts a polypeptide of 129 amino acids, preceded by a signal peptide of 21 amino acids. The predicted polypeptide is unique in its extremely high content of methionine (22.5%). The maize inbred line BSSS-53, which has increased seed methionine due to overproduction of this protein, was compared to W23, a standard inbred line. Northern blot analysis showed that the relative RNA levels for the 10 kDa zein were enhanced in developing seeds of BSSS-53, providing a molecular basis for the overproduction of the protein. Southern blot analysis indicated that there are one or two 10 kDa zein genes in the maize genome.

Key words: Zein – *Zea mays* – Gene expression – Seed development – High methionine protein

Introduction

The expression of seed storage protein genes is tissue-specific and developmentally regulated. These genes are expressed only during defined stages of seed development, and the expression is limited to the embryo and/or endosperm tissue of developing seeds. In agriculturally important seed crops the expression of storage protein genes directly affects the nutritional quality of the seed protein. In maize (*Zea mays* L.) the prolamine (zein) fraction of storage proteins comprises over 50% of the total protein in the mature seed. Zein polypeptides contain extremely low levels of the essential amino acids lysine, tryptophan and, to a lesser extent, methionine. Maize seed protein is deficient in these amino acids because such a large percentage of the total protein is contributed by the zeins. Several mutations in maize affect the expression of zein genes and result in improved nutritional quality of the seed protein. For example, in the seeds of plants homozygous for the recessive mutation

opaque-2 (o2) (Mertz et al. 1964), the levels of the M_r 22000 (22 kDa) zeins are drastically reduced (Misra et al. 1975; Soave et al. 1976). There is a concomitant increase in the proportion of more nutritionally balanced proteins deposited in the seed. The net result is an increase in the levels of lysine and tryptophan in the seed (Misra et al. 1972).

The inbred line BSSS-53 was characterized by a seed methionine content 30% higher than that of other inbred lines tested (Phillips et al. 1981). It was later shown (Phillips and McClure 1985) that the increased methionine content in BSSS-53 seeds was the result of a twofold increase in the level of the methionine-rich 10 kDa zein storage protein fraction. The other zein subfractions were present in levels comparable to those found in other inbred lines, and the total protein content and kernel phenotype were normal. Amino acid analysis indicated that the 10 kDa zein fraction was composed of approximately 20% methionine.

We are investigating the differential expression of the 10 kDa zein in BSSS-53 compared to other maize strains. Due to the high methionine content of the 10 kDa zein, and since methionine is specified by a unique triplet codon (ATG), the following approach was taken to isolate a cDNA clone encoding this polypeptide. A 10 kDa zein polypeptide was isolated, and the sequence of the N-terminal 30 amino acids was determined. Based on the amino acid sequence, two mixed oligonucleotides were synthesized and used to screen a maize endosperm cDNA library. A full-length cDNA clone encoding the 10 kDa zein was isolated by this procedure. We report here the purification and N-terminal amino acid sequence of the 10 kDa zein polypeptide, and the nucleotide sequence of the cDNA clone encoding this protein.

The 10 kDa zein is distinguished by its extremely high methionine content (22.5%). The increased expression of the 10 kDa zein protein in BSSS-53 was found to be correlated with elevated levels of 10 kDa zein RNA in the endosperm of developing seeds. Southern blot analysis of maize genomic DNA indicated that the 10 kDa zein subfraction is encoded by one or two structural genes.

Materials and methods

Plant material. Seeds of maize (*Z. mays* L.) inbred lines W64A, W23 and BSSS-53 were kindly provided by R.L. Phillips, Dept. of Agronomy, University of Minnesota, St. Paul, MN 55108, USA. Endosperm samples were obtained from seeds of hand-pollinated plants grown in the field

* Present address: Kallestad, A Division of Eriamont, 1000 Lake Hazeltine Dr., Chaska, MN 55318, USA

Offprint requests to: J.W. Messing

in 1986. Leaf samples were obtained from seedlings grown in a growth chamber.

Protein extraction. Zm protein fractions were isolated as described by Phillips and McClure (1985). Protein concentrations were determined against a bovine serum albumin standard curve according to the method of Peterson (1977).

SDS-polyacrylamide gel electrophoresis and isoelectric focusing. SDS polyacrylamide gel electrophoresis (SDS-PAGE) was carried out according to the method of Laemmli (1970). Separating gels of 15% acrylamide were 110 × 140 mm; preparative gels were 3 mm thick and analytical gels were 1.5 mm thick. Proteins were visualized after preparative SDS-PAGE by soaking the gel in 0.25 M KCl, 1 mM dithiothreitol as described by Hager and Burgess (1980). Analytical SDS-PAGE gels were stained with Coomassie blue.

Isoelectric focusing (IEF) was performed on 2 mm slab gels using an LKB Multiphor apparatus. IEF gels were 5% acrylamide, 0.4 M urea, and contained 2% pH 5-8 ampholytes (Serva). IEF gels were run at 12 W constant power for 2 h at 10°C and then at 15 W for 30 min at the same temperature. Proteins were visualized after IEF by soaking the gel in 10% trichloroacetic acid (TCA) or by Coomassie staining. For preparative IEF, only a portion of the gel was treated with 10% TCA to enable localization of the protein bands within the remainder of the gel.

Elution of proteins from polyacrylamide gels. SDS-PAGE or IEF gel slices that contained the protein bands of interest were minced and covered with SDS-gel electrophoresis buffer. The protein was then electroeluted from the gel pieces. Eluted protein was dialyzed extensively against 70% ethanol and lyophilized. Protein samples were further purified by reverse phase HPLC (Mahoney and Hermanson 1980).

Amino acid analysis. Samples were hydrolyzed at 110°C in sealed, evacuated tubes with glass-distilled, 6 N HCl for 24 h. The protein was not reduced or alkylated. Analyses were carried out on a Beckman System 6300 amino acid analyzer.

Amino acid sequence analysis. Samples were degraded in a Beckman Model 890 D sequencer according to the procedure of Edman and Begg (1967) using a slight modification of the Beckman 0.1 M Quadrol peptide program (No. 345801). Prior to the addition of the sample to the sequencer cup, 2 mg of Polybrene were dissolved in 0.7 ml of 50% acetic acid and applied to the cup of the sequencer, dried under vacuum, and subjected to three complete cycles of automated Edman degradation. The material under investigation was then introduced into the cup and sequentially degraded. Products generated by the sequencer were converted to their phenylthiohydantoin (Pth) derivatives as previously described (Mahoney and Nute 1980). Pth-amino acid derivatives were identified using a Varian Model 5560 ternary high performance liquid chromatograph, equipped with a Varian Model 8000 autosampler modified for reduced sample loss, a Hewlett-Packard 3390 recording integrator and a Beckman 0.46 × 25 cm Ultrasphere ODS 5 µm column (Zimmerman et al. 1977; Nute and Mahoney 1980). Pth-amino acids were identified based upon comparison with known standards. When the signal to noise ratio fell

below 2, identifications were not attempted. Stepwise yields for the degradation ranged from 92% to 96% and only one sequence was observed.

cDNA library construction. Poly(A)⁺ RNA enriched for zea encoding sequence was prepared from sucrose gradient purified zea endosperm isolated from endosperm of maize inbred W22 at 22 days post pollination (Burr and Burr 1976).

The synthesis of cDNA was carried out by a vector-prime method designed for use with advanced pUC plasmids (J. Hunsperger and J. Rubenstein, in preparation). The vector used was pUC119 (Vierra and Messing 1987). Vector DNA was digested with *Kpn*I and 3-tailed. The DNA was then digested with *Bam*HI to provide a single priming site for reverse transcriptase. Ten micrograms of methyl mercury denatured poly(A)⁺ RNA was annealed to 2 µg of vector primer in a first strand synthesis reaction polymerized by AMV-L⁺ reverse transcriptase. Following second strand synthesis (Okajima and Berg 1982), duplex cDNA-vector was methylated with *Eco*RI methylase, ligated to *Eco*RI linkers, and digested with *Eco*RI. The entire population of linear cDNA-vector species was size fractionated on agarose gels, using the method of Hanahan (1983). Circularized cDNA-vector DNA from the individual fractions was used to transform a DH5 (Hanahan 1983) derivative bearing *F⁺ lac P⁺ Z⁺ Tn5 Y⁺ A⁺*. The resulting maize endosperm protein body cDNA library designated PB-2, consisted of 4.2×10^6 independent clones.

Screening the cDNA library. Colony hybridization using synthetic oligonucleotide probes was performed according to the protocol of Woods (1984). Only those colonies showing hybridization on duplicate replica filters were chosen for further analysis. Positive colonies were picked and colony-purified. Positive clones were verified by hybridization of the oligonucleotide probes to Southern blots of restriction enzyme-digested plasmid DNAs.

Template preparation, deletion subcloning, and DNA sequencing. Single-stranded plasmid DNA for deletion subcloning and DNA sequencing was prepared as previously described (Vierra and Messing 1987). A set of overlapping sequential deletion subclones for DNA sequencing was prepared as described by Dale et al. (1987). DNA sequencing was carried out by the dideoxy method (Sanger et al. 1977) with [γ -³²S]dATP using the protocol outlined in a kit purchased from Amersham. All templates were sequenced at least twice and the sequence was determined on both DNA strands.

Maize DNA and RNA isolation. Genomic DNA was isolated from leaf tissue of 3-week old maize seedlings as described by Shure et al. (1983).

For RNA isolations, endosperms were dissected from maize kernels harvested at specific times after pollination. The endosperms were frozen in liquid nitrogen and stored at -80°C until needed. Endosperm samples (0.5 g) were ground to a fine powder in liquid nitrogen, and total RNA was isolated as described by Berry et al. (1985).

Southern blot and northern blot analysis. Maize genomic DNA samples were digested with restriction enzymes and fractionated on 0.8% agarose gels. After staining and phos-

ography, the DNA was partially depurinated (Wahl et al. 1979) and transferred to Nytran membrane (Schleier and Schuell) according to Southern (1975). Filters were prehybridized and hybridized according to the manufacturer's specifications. Hybridized filters were washed twice at room temperature for 15 min in $6 \times$ SSC, 0.1% SDS, 0.05% sodium pyrophosphate (NaPPi), then twice at 37°C for 15 min in $1 \times$ SSC, 0.5% SDS, 0.05% NaPPi. The final stringent wash was for 60 min at 65°C in $0.1 \times$ SSC, 1% SDS, 0.05% NaPPi.

Northern blot analysis of maize endosperm total RNA was carried out on 1.2% agarose-formaldehyde gels. Denaturation, electrophoresis and transfer of RNA were performed as described by Mamatis et al. (1982), except that Nytran membrane was used in place of nitrocellulose. Filters were prehybridized and hybridized according to the manufacturer's specifications. Hybridized filters were washed as described above for Southern blot hybridization.

The DNA probe used for Southern and Northern blot analysis was a deletion subclone of 10kZ-1 (see Fig. 4), designated 10kZ-1J43. 10kZ-1J43 lacks the entire poly-A tail and approximately 50 nucleotides 5' to the poly-A tail of 10kZ-1. 10kZ-1J43 DNA was labeled with [$\gamma^{32}\text{P}$]dCTP (New England Nuclear, 800 Ci/mmol) by nick translation (Rigby et al. 1977). Average specific activity of the labeled probes was 1×10^6 cpm/ μg . Hybridized filters were exposed to Kodak XAR-5 X-ray film for 1–72 h at -80°C with a Dupont Cronex intensifying screen.

Results

Protein purification

Zein-1 and zein-2 fractions were isolated from kernels of the maize inbred lines W64A and BSSS-53 as described in Materials and methods. SDS-PAGE analysis of the zein-2 fractions from W64A and BSSS-53 demonstrated that the 10 kDa zein was present in higher proportion in BSSS-53 (Fig. 1, compare lanes 3 and 4). The 10 kDa zein subfraction was isolated from these two inbred lines by preparative SDS-PAGE (lanes 5 and 6). The 10 kDa zein fractions isolated from the two inbred lines were similar in amino acid content (Table 1). When N-terminal amino acid sequencing was attempted on SDS-PAGE-purified 10 kDa zein from BSSS-53, it was found that this fraction was heterogeneous, and no single N-terminal sequence was obtained. The 10 kDa zein was then fractionated by isoelectric focusing (Fig. 2) and indeed several components were detected. The major IEF band was purified by preparative IEF in polyacrylamide slab gels. The purified polypeptide is shown in lanes 7 and 5 of Figs 1 and 2, respectively. We were able to obtain a partial N-terminal amino acid sequence of this protein fraction (Fig. 3A).

Amino acid sequence analysis

The 10 kDa zein protein presented problems due to its lack of solubility in aqueous buffers. Attempts at reduction and S-pyridylethylation met with extremely low yields (as determined by amino acid analysis), with commensurate loss of material. As such, amino acid sequencing was done in the absence of reduction and alkylation, knowing that this would not allow the identification of cysteine. As shown in Fig. 3 and Table 2, we were able to order the first 30 amino acids, with 5 of the identifications in question. The

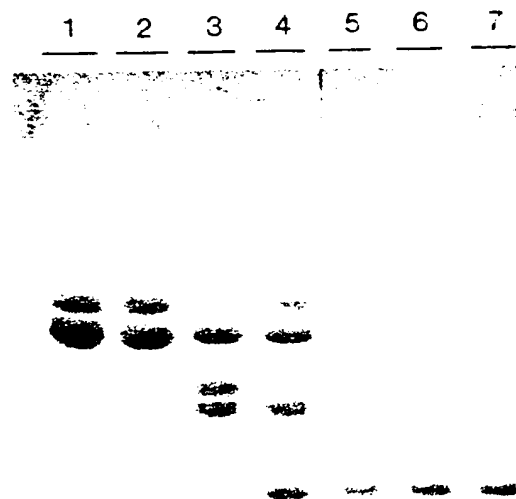


Fig. 1. SDS-polyacrylamide gel electrophoresis (SDS-PAGE) analysis of zein polypeptides. Zein fractions (20–100 μg) isolated from seeds of the inbred lines W64A and BSSS-53 were separated by SDS-PAGE on a 5% gel, stained with Coomassie. Lanes 2 and 3, zein-2 fractions from W64A and BSSS-53, respectively; lanes 4 and 5, zein-1 fractions from W64A and BSSS-53, respectively; lanes 6 and 7, SDS-PAGE-purified 10 kDa zein from W64A and BSSS-53, respectively; lane 7, 10 kDa zein from BSSS-53 purified by isoelectric focusing.

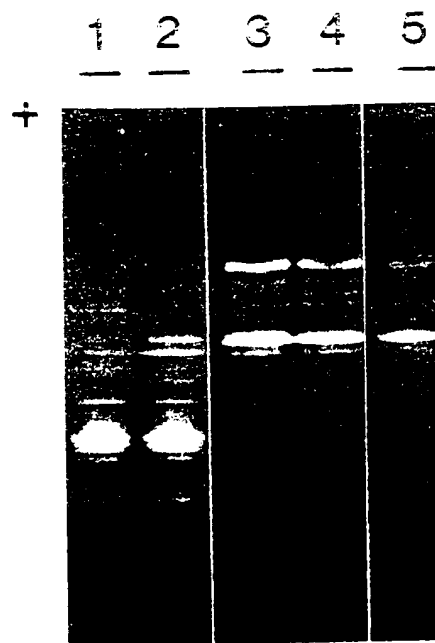


Fig. 2. Analytical isoelectric focusing (IEF) gel of zein polypeptides. Zein fractions (20–100 μg) isolated from the inbred lines W64A and BSSS-53 were analyzed by IEF. Proteins were visualized by soaking the gel in 10% TCA, and the gel was photographed on a dark background with side lighting. Lanes 2 and 3, zein-2 fractions from W64A and BSSS-53, respectively; lanes 4 and 5, 10 kDa zein fractions from W64A and BSSS-53, respectively; purified by SDS-polyacrylamide gel electrophoresis; lane 5, IEF-purified 10 kDa zein from BSSS-53. The anode was at the top.

Table 1. Amino acid compositions of 10 kDa zein polypeptide fractions isolated from the inbred lines W64A and BSSS-53, and amino acid composition of 10kZ-1 derived from the nucleotide sequence

Amino acid	mol/100 mol		
	W64A	BSSS-53	10kZ-1 cDNA
Asx*	3.4	3.7	3.1
Asn	—	—	2.3
Asp	—	—	0.8
Thr	4.0	4.5	3.9
Ser	5.8	6.5	6.2
Glx*	14.6	13.8	13.6
Gln	—	—	11.6
Glu	—	—	0.0
Pro	15.4	14.6	17.5
Gly	3.5	5.7	3.1
Ala	6.6	5.7	5.4
Cys	—	—	3.9
Val	4.2	4.2	3.9
Met	18.4	17.8	22.5
Ile	2.6	2.6	2.7
Leu	13.1	12.0	11.6
Tyr	1.2	1.0	0.8
Phe	4.9	4.8	3.9
His	2.3	2.5	2.3
Lys	0.1	0.5	0.0
Arg	0.0	0.0	0.0
Trp	—	—	0.0

* Asx and Glx refer to (Asp + Asn) and (Glu + Gln), respectively. values for Cys and Trp were not determined in amino acid analysis of polypeptide fractions

amino terminal residue was identified as threonine in initial sequence analyses, and as glutamine in a subsequent analysis; however, this was the only disagreement in the data. Two residues had more than a single unit residue: residue 12 had both asparagine and proline, and residue 21 had both glutamine and methionine (Fig. 3 and Table 2). In the identification of threonine at residue 23, although the yield was poor, there appeared to be a small amount of dehydro-threonine present.

The derived nucleotide sequence for the region between amino acid residues 20 and 26 was chosen for the synthesis of two mixed oligonucleotide probes of 20 nucleotides in length (Fig. 3B). One of the probes (probe M) reflected the methionine at residue 21, while the second probe (probe G) reflected the glutamine at this position. The oligonucleotides were designed to be complementary to the mRNA and therefore to the coding strand of the DNA so that positive clones could be quickly verified by DNA sequencing using the oligonucleotide probes as sequencing primers. The oligonucleotide probes were specific for the 10 kDa zein since the region chosen contained 3 (probe G) or 4 (probe M) methionine residues. With the exception of the 15 kDa zein, methionine is a rare (1%–2%) amino acid in other zeins. The mature 15 kDa zein contains 18 methionine residues (Marks et al. 1985b; Pedersen et al. 1986) but has no homology to the oligonucleotides.

Screening the cDNA library

Southern blot analysis of plasmid DNA isolated from 6 of the 8 size fractions of the cDNA library indicated that

1	5	10
Thr		
Gln - His - Ile - Pro - Gln - His - Leu - Pro - (Thr - Val		
11	15	20
Met - <u>Asn</u> - <u>Leu</u> - Gly - Thr - Met - Asn - (Tyr) - (Pro) - Met		
Pro		
21	25	30
Gln		
Met - Tyr - Thr - Met - Met - Gln - Gln - (Gly) - (Leu) - Ala		
Met		

A

Probe M:

Met	Met	Tyr	Thr	Met	Met	Gln	Prot. seq.
5' ATG	ATG	TAPy	ACN	ATG	ATG	CAPu	Coding strand
TAC	TAC	ATPu	TGN	TAC	TAC	GT 5'	Oligo. seq.

Probe G:

Met	Gln	Tyr	Thr	Met	Met	Gln	Prot. seq.
5' ATG	CAPu	TAPy	ACN	ATC	ATG	CAPu	Coding strand
TAC	GTPy	ATPu	TGN	TAC	TAC	GT 5'	Oligo. seq.

B

Fig. 3 A, B. A, N-terminal amino acid sequence of the 10 kDa zein polypeptide fraction from BSSS-53 purified by isoelectric focusing. Underlined residues indicate that stretch of amino acid residues that was used for the prediction-guided synthesis of oligonucleotide probes. Parentheses indicate amino acid residues only tentatively identified. Sequence microheterogeneity was detected at residues 12 and 21. B Sequence of the two synthetic mixed oligonucleotide probes, as derived from the amino acid sequence. In, purine; Py, pyrimidine; N, any base.

fraction 6 contained most of the sequences hybridizing to the oligonucleotide probes (data not shown). Approximately 20000 colonies from fraction 6 were screened by colony hybridization to the 2 mixed oligonucleotide probes. Approximately 200 colonies showed strong hybridization to probe G after washing the filters at 37°C. Probe M hybridized to the majority of the same colonies after washing the filters at 25°C. However, no hybridization above background was detected with probe M after the filters were washed at 37°C. Therefore, only positive colonies detected with probe G were chosen for further analysis. Using probe G as a sequencing primer, we were able to identify a clone which encoded a polypeptide with the same amino acid sequence as the 10 kDa zein. This clone, designated 10kZ-1, was chosen for complete DNA sequence determination.

Nucleotide sequence of 10kZ-1

The DNA sequence of 10kZ-1 is shown in Fig. 4. This cDNA clone encodes a polypeptide of 129 amino acids preceded by a leader peptide of 21 amino acids. The site of cleavage of the leader peptide was determined by comparing the amino acid sequence of the mature 10 kDa zein protein with the derived amino acid sequence of the cDNA clone. As expected from the results of the colony hybridization, the cDNA clone encodes a polypeptide with a glutamine rather than a methionine at residue 21 of the mature polypeptide. The cDNA clone has 21 nucleotides 5' to the ATG and 96 nucleotides 3' to the TAG stop codon. There is a consensus poly(A) addition signal (AATAAA) 25 nucleotides 5' to the poly(A) tail, similar to other eukaryotic genes

Table 2. Yields and identification of the products generated by automated Edman degradation

Position	Amino acid	Yield (nmole)
1	Gln	26.1
2	His	22.2
3	His	30.8
4	Pro	32.3
5	Gly	28.5
6	His	19.0
7	Leu	25.3
8	Pro	18.7
9	(Thr)	9.2
10	Val	22.1
11	Met	20.1
12	Asn-Pro	7.4-7.5
13	Leu	21.1
14	Gly	12.4
15	Thr	8.3
16	Met	17.0
17	Asn	15.3
18	(Gly)	11.6
19	(Ser-Ala-Pro-Arg)	trace amounts
20	Met	5.9
21	Gln-Met	4.1-9.1
22	Thr	5.0
23	(Thr)	4.8
24	Met	7.9
25	Met	10.8
26	Gln	2.1
27	Gln	2.8
28	(Gly)	1.0
29	(Leu)	4.5
30	(Ala)	3.3

* Only 75% of each product generated by the sequencer was analyzed. Yields listed above were normalized to 100% injection.

(Nevins 1983). The amino acid composition of the mature polypeptide encoded by the cDNA clone agrees with the amino acid analysis of the 10 kDa zein proteins (Table 1).

It is interesting to note that the DNA sequence of this clone differs from that predicted by the protein sequence. At amino acid position 23, the cDNA clone encodes a cysteine, while a threonine residue was identified in the N-terminal amino acid sequence. This discrepancy may represent an allelic difference, since the protein was isolated from the inbred line BSSS-53, while the cDNA library was prepared from poly(A)⁺ RNA from W23. Alternatively, this residue might represent an additional sequence microheterogeneity which went undetected (as discussed earlier, the protein was not derivatized prior to amino acid sequence analysis, and cysteine could not be identified). With the exception of amino acids that were only tentatively identified, the remainder of the predicted amino acid sequence agreed precisely with the N-terminal amino acid sequence.

Developmental expression of the 10 kDa zein

It had been shown that the level of 10 kDa zein protein was higher in BSSS-53 seeds than in W23 seeds (Phillips and McClure 1985). To determine whether the differential accumulation of the 10 kDa zein protein in mature kernels of BSSS-53 and W23 was correlated with differential levels of 10 kDa zein RNA in the developing endosperm, we analyzed RNA from the progeny of self-pollinated W23 and

10	30	50
GGAAGCAAGGACACACCCGCGCATGGCAGCCAAGATGCTTGCATTGTCCTCTCTAGCT		
	MetAlaAlaAlaAlaMetLeuAlaLeuPheAlaLeuLeuAla	
70	90	110
CTTTGTGCAAGCGCCACTAGTGGCAGCCATATTCAGGGGCACTGGCCACCACTCATGCCA		
LeuCysAlaSerAlaThrSerAlaThrHisIleProGlyHisLeuProProValMetPro		
130	150	170
TTGGGTACCATGAAGCCATGCATGCAGTACTGCATGATCCACAGGGGCTTGGCAGCTTC		
LeuGlyThrMetAsnProCysMetGlnTyrCysMetMetGlnGlnLeuLeuAlaLeuAlaSerLeu		
190	210	230
ATGGCGTGTCCGTCCTGATGCTGCAGCAACTGTGGGCTTACCGCTTCAGACGATGCCA		
MetAlaCysProSerLeuMetLeuGlnGlnLeuLeuAlaLeuProLeuGlnThrMetPro		
250	270	290
GTGATGATGCCACAGATGATGACGCCCTAACATGATGTCAACATTGATGATGCCAGCATG		
ValMetMetProGlnMetMetThrProAsnMetMetSerProLeuMetMetProSerMet		
310	330	350
ATGTCACCAATGGTCTTGGCGAGCATGATGTCGCAATGATGATGCCCAATGTCACTGC		
MetSerProMetValLeuProSerMetMetSerGlnMetMetMetProGlnCysHisCys		
370	390	410
GACGCGCTCTCGCAGATTATGCTGCAACAGCAGTTACCATTCATGTTCAACCCATGGCC		
AspAlaValSerGlnIleMetLeuGlnGlnLeuProPheMetPheAsnProMetAla		
430	450	470
ATGACGATTCCACCCATGTTCTTACAGCAACCGTTGTTGCTCCTGCATTCTAGATAGAA		
MetThrIleProProMetPheLeuGlnGlnProPheValGlyAlaAlaPhe		
490	510	530
ATATTGTGTTGTACCGAATAATGAGTTGACATGCCATCGCGTGTGATTCATTATTAAGL		
550	570	
ATAAAACAAGTTTCTCTTATTATCTTTT(A) _n		

Fig. 4. Nucleotide sequence and derived protein sequence of 10kZ-1. The arrow indicates the N-terminal amino acid of the mature polypeptide, as determined by N-terminal amino acid sequencing. The sequence upstream of the arrow encodes a 21-amino acid signal peptide. The consensus poly(A)⁺ addition sequence (ATAAA)_n is underlined.

BSSS-53 plants. Total RNA was prepared from endosperm tissue isolated at 5 time points post-pollination. The RNA samples were compared by Northern blot analysis using the 10 kDa zein probe described in Materials and methods. As shown in Fig. 5, 10 kDa zein transcripts were first detected at 12 days post-pollination. The level of 10 kDa zein transcripts reached a peak at 15-18 days post-pollination and declined after that point. This pattern of developmental expression is similar to the results obtained for other zein genes (Marks et al. 1985a), i.e., zein transcripts were first observed at approximately 12 days post-pollination, their levels peaked between 18 and 21 days post-pollination and declined slowly after that time. As shown in Fig. 5, 10 kDa zein RNA levels were significantly higher in BSSS-53 than in W23 at all time points analyzed.

Estimate of the 10 kDa zein gene copy number

A possible mechanism for the elevated 10 kDa zein RNA levels in seeds of BSSS-53 is through amplification of the 10 kDa zein structural genes. Therefore, we compared the genomic DNAs of BSSS-53 and W23 by Southern blot hybridization. Genomic DNA was isolated from seedlings of BSSS-53, W23 and the cross W23 × BSSS-53. The DNA samples were analyzed by Southern blot hybridization using the 10 kDa zein probe (Fig. 6). Comparison of the intensity of hybridization of the probe to genomic DNA versus the gene-copy reconstruction, indicated that the 10 kDa zein gene was present in only one or two copies in both W23 and BSSS-53. This result demonstrated that there was no gross amplification of the 10 kDa zein genes in BSSS-53. The results presented in Fig. 6 also demonstrate the existence of restriction fragment length polymorphisms

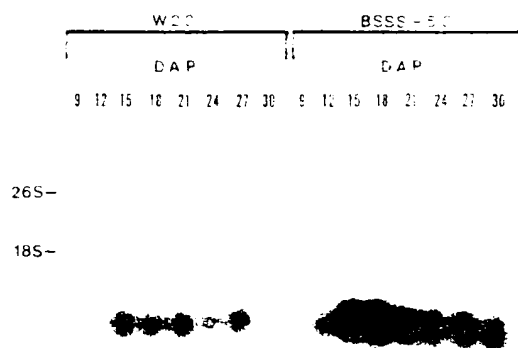


Fig. 5. Northern blot analysis of maize endosperm total RNA from W23 and BSSS-53. Total RNA (5 µg) isolated from endosperms harvested at 9, 12, 15, 18, 21, 24, 27 and 30 days after pollination (DAP) was denatured, separated on a 1.2% agarose-formaldehyde gel, transferred to Nytran membrane and probed with nick-translated 10KZ-1.43 DNA. The positions of the maize 18S and 26S rRNAs are indicated on the left.

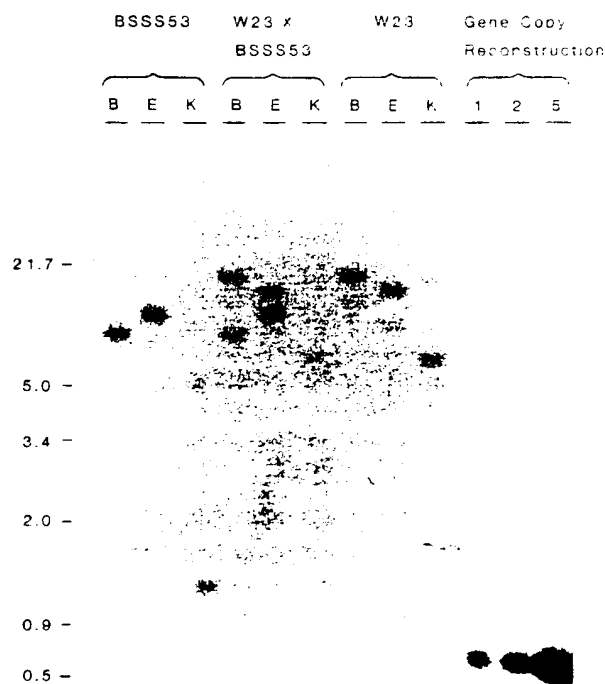


Fig. 6. Southern blot of genomic DNA from BSSS-53, W23 and the F1 W23 x BSSS-53. Samples (3 µg) of genomic DNA were digested with *Bam*HI (B), *Eco*RI (E), or *Kpn*I (K), size fractionated on a 0.8% agarose gel, transferred to Nytran membrane and probed with nick-translated 10KZ-1.43 DNA. Samples of *Eco*RI-digested 10KZ-1 DNA were diluted to 1, 2 and 5 gene equivalents and loaded on the same gel as concentration standards. The numbers on the left indicate the positions of size standards (kb).

(RELPs) between the DNA from W23 and BSSS-53. DNA from the F1 parent lines are used here to detect polymorphisms. These RELPs will be useful for discriminating between the 19 kDa zein genes from W23 and BSSS-53 in future experiments. An additional weakly-hybridizing band is routinely observed on genomic southern blots. This band may represent a divergent 19 kDa zein gene. Further studies are being conducted to investigate this possibility.

Discussion

The primary storage proteins in the maize seed are a group of alcohol-soluble polypeptides called zeins. Collectively, the zeins account for over 50% of the protein content in a mature maize kernel (Wilson 1983). When subjected to SDS-PAGE, zein polypeptides separate into 5 subclasses with apparent molecular weights of 27000, 22000, 19000, 15000 and 10000 (Granazza et al. 1977). The 22 kDa and 19 kDa zein polypeptides represent the majority (75%–80%) of the zein fraction and are extracted with 70% ethanol (the zein-1 fraction). When a reducing agent, such as 6-mercaptoethanol, is present, additional polypeptides of 27 kDa, 15 kDa and 10 kDa are extracted. This latter group of polypeptides has been referred to as alcohol-soluble reduced glutelin (Paulis and Wall 1971), or zein-2 (Sodek and Wilson 1971). The 22 kDa and 19 kDa zeins are encoded by a complex multi-gene family with a pool of active and inactive genes (reviewed in Heidecker and Messing 1986). In contrast, the 15 kDa and 27 kDa zeins are each encoded by only one or two genes (Wilson and Larkins 1984; Das and Messing 1987).

The analysis of gene copy number is supported by isoelectric focusing analysis and two-dimensional gel electrophoresis of zein polypeptides. While the zein-1 polypeptides show extensive charge heterogeneity (Fagnetti et al. 1977; Hagen and Rubenstein 1980; Hurkman et al. 1981), it has been reported that the 27 kDa, 15 kDa and 10 kDa zeins are each represented by polypeptides of a single isoelectric point (Hurkman et al. 1981; Marks et al. 1985a). The SDS-PAGE-purified 10 kDa zein produced multiple bands on IEF gels (Fig. 2). At present, it is not known whether the additional bands represented additional 10 kDa zein proteins, or whether they were artifacts of the purification process. The microheterogeneity detected in the N-terminal amino acid sequence suggests that the 10 kDa zein subclass contains many very similar polypeptides. However, since positive colonies were only detected with probe G, it is unclear at this time whether or not the glutamine versus methionine at residue 21 represents an allelic variation.

Zein polypeptides are characterized by their high content of proline, glutamine, leucine and alanine (Granazza et al. 1977; Wilson 1983). The 27 kDa, 15 kDa and 10 kDa zeins are distinguished from the 22 kDa and 19 kDa classes by their increased content of cysteine and methionine (Granazza et al. 1977; Eisen et al. 1981). It has been proposed (Paulis et al. 1969) that these polypeptides interact through intermolecular disulfide bonds, which results in their efficient extraction only under reducing conditions. The 10 kDa zein is remarkable for its extremely high methionine content (22.5%). With the exception of the 15 kDa zein, where methionine constitutes approximately 10% of the amino acids (Marks et al. 1985b; Pedersen et al. 1986), methionine is a rare (1%–2%) amino acid in other zein

polypeptides (Granazza et al. 1977; Wilson 1983), and other proteins in general. In total, the sulfur-containing amino acids comprise over 25% of the amino acids in the 10 kDa zein.

In the maize kernel, zein polypeptides are found sequestered in membrane-bound granules called protein bodies (Wolf et al. 1967). The deposition of zein polypeptides into protein bodies is believed to occur via cotranslational transport into the rough endoplasmic reticulum (Larkins and Hurkman 1978; Burr and Burr 1981). The 10 kDa zein cDNA clone encodes a polypeptide which is 21 amino acids longer at the N-terminus than the mature polypeptide. The sequence of the N-terminal 21 amino acids shows striking homology to the signal peptides of other zeins (Messing 1987). Therefore, we believe that this sequence constitutes a signal peptide, and it is likely that the 10 kDa zein is deposited into protein bodies in the endosperm.

The level of the 10 kDa zein protein was previously shown to be higher in seeds of BSSS-53 than in seeds of W23 (Phillips and McCure 1985). This difference was correlated with different levels of 10 kDa zein RNA in developing endosperms from these two inbred lines (Fig. 5). At all time points analyzed, 10 kDa zein transcripts were more abundant in BSSS-53 as compared to W23, while the overall developmental profile appeared to be unaltered. Quantitative data indicate that 10 kDa zein RNA levels are 2- to 5-fold higher in BSSS-53 than in W23, depending on the developmental time point (I. Kirihara and J. Messing, in preparation). The increased 10 kDa zein RNA levels may be due to increased transcription of the 10 kDa zein gene(s) in BSSS-53, or possibly to a difference in stability of 10 kDa zein transcripts between the two inbred lines. Regardless of the cause however, it is likely that the increased level of 10 kDa zein RNA contributes to the increased level of 10 kDa zein protein found in the mature seed.

The increased expression of the 10 kDa zein in BSSS-53 represents an interesting example of differential gene expression. While mutations such as *opaque-2* (Misra et al. 1972) and *floury-2* (Nelson et al. 1965; Haisel et al. 1973) result in a decrease in zein proteins in the seed, in BSSS-53 seeds a subclass of zein proteins is increased. In *opaque-2* mutants, 22 kDa zein mRNA and protein levels are drastically reduced (Misra et al. 1975; Soave et al. 1976; Pedersen et al. 1980; Burr and Burr 1982). The *opaque-2* mutation is located on maize chromosome 7, unlinked to some of the zein genes whose expression it affects (Soave et al. 1978). The *opaque-2* gene is thought to be a regulatory gene involved in zein gene expression. The genetic element responsible for overexpression of the 10 kDa zein protein is located on chromosome 4 (Benner and Phillips 1986). Recently, it has been determined that this element is not linked to the 10 kDa zein structural gene(s) in BSSS-53 (M. Benner and R. Phillips, personal communication). Since the element responsible for the overexpression is not linked to the structural gene(s), it may represent a regulatory gene which enhances the expression of the structural gene(s). In contrast to *opaque-2*, which affects the expression of a large family of genes, molecular analysis of the overexpression of the 10 kDa zein may be simplified due to the small number of 10 kDa zein genes.

Acknowledgment. The authors gratefully acknowledge Elizabeth D. Lewis for the preparation of protein body-bound poly A⁺ RNA, Immuno-Suclear Corporation, Stillwater, MN for providing their facilities for protein sequencing and Dr. David Mace for the

preparation of synthetic oligonucleotides. This work has been supported by the U.S. Department of Energy, grant no. DE-AC05-84OR21400.

References

- Benner M, Phillips RL (1986) Chromosomal location of a gene controlling arg-methionine zein synthesis. *Maize Genetics Cooperation New Letter* 66:114
- Burr JG, Nikolau BJ, Carr JE, Klassig DJ (1987) Transcriptional and post-transcriptional regulation of ribulose-1,2-bisphosphate carboxylase gene expression in light- and dark-grown amaranth cotyledons. *Mol Cell Biol* 7:2238-2246
- Burr B, Burr EA (1976) Zein synthesis in maize by polyribosomes attached to protein bodies. *Proc Natl Acad Sci USA* 73:515-519
- Burr EA, Burr B (1981) *In vitro* uptake and processing of prozein and other maize proprotein by maize membranes. *J Cell Biol* 90:427-432
- Burr EA, Burr B (1982) Three mutations in *Zea mays* affecting zein accumulation: a comparison of zein polypeptides, amino synthesis and processing, mRNA levels, and genomic organization. *J Cell Biol* 94:291-296
- Dale PMK, McCure BA, Froughins JP (1985) A rapid, single stranded cloning strategy for producing a sequence specific overlapping clones for use in DNA sequencing: application to sequencing the corn mitochondrial 16S rDNA. *Plasmid* 13:31-40
- Das OP, Messing J (1987) Allelic variation and differential expression at the 27 kDa zein locus in maize. *Mol Cell Biol* 7:4496-4497
- Edman P, Begg G (1967) A protein sequencer. *Eur J Biochem* 1:50-51
- Esen A, Bietz JA, Parry JW, Wall GS (1981) Fractionation of alcohol-soluble reduced corn glutens on phosphocellulose and partial characterization of two proline-rich fractions. *Cereal Chem* 58:534-537
- Granazza E, Vighienzi V, Righetti PG, Salamini F, Soave G (1977) Amino acid composition of zein molecular components. *Phytochemistry* 16:315-317
- Hagen G, Rubenstein J (1980) Two-dimensional gel analysis of the zein proteins in maize. *Plant Sci Lett* 19:217-223
- Hager DA, Burgess RF (1980) Elution of proteins from sodium dodecyl sulfate-polyacrylamide gels, removal of sodium dodecyl sulfate, and renaturation of enzymatic activity. Results with sigma subunit of *Escherichia coli* RNA polymerase, a heat-labile DNA topoisomerase, and other enzymes. *Anal Biochem* 109:76-80
- Handal AK (1985) Studies on transformation of *Escherichia coli* with plasmids. *J Mol Biol* 166:557-580
- Haisel LW, Fan CY, Nelson GE (1973) The effect of the *opaque-2* gene on the distribution of protein fractions and distribution of zein in endosperm. *Cereal Chem* 50:383-392
- Hendricks G, Messing J (1986) Structural analysis of maize genes. *Ann Rev Plant Physiol* 37:439-460
- Hurkman WJ, Smith LD, Benner J, Larkins BA (1978) Subcellular compartmentalization of maize storage proteins in xenopus oocytes injected with zein messenger RNA. *J Cell Biol* 89:292-299
- Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* 227:680-685
- Larkins BA, Hurkman WJ (1978) Synthesis and deposition of zein in protein bodies of maize endosperm. *Plant Physiol* 62:256-263
- Manthey WC, Hermanson MA (1980) Separation of large denatured peptides by reverse phase high performance liquid chromatography. Trifluoroacetic acid is a peptide solvent. *J Biol Chem* 255:11199-11203
- Manthey WC, Nute PL (1980) Fetal hemoglobin of the Rhesus monkey, *Macaca mulatta*. Complete primary structure of the γ chain. *Biochemistry* 19:4436-4442

- Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor: Laboratory Press, New York
- Marks MD, Lindell JS, Larkins BA (1985a) Quantitative analysis of the accumulation of zein mRNAs during maize endosperm development. *J Biol Chem* 260:16445-16450
- Marks MD, Lindell JS, Larkins BA (1985b) Nucleotide sequence analysis of zein mRNAs from maize endosperm. *J Biol Chem* 260:16451-16459
- Mertz ET, Bates LS, Nelson OE (1964) Mutant gene that changes protein composition and increases lysine content of maize endosperm. *Science* 145:279-280
- Messing J (1987) The genes encoding seed storage proteins in higher plants. In: Rigny P (ed) Genetic Engineering, vol 6. Academic Press, London: 2-46
- Misra PS, Ambunathan R, Mertz ET, Glover DV, Barbosa HML, McWhirter KS (1972) Endosperm protein synthesis in maize mutants with increased lysine content. *Science* 176:1425-1426
- Misra PS, Mertz ET, Glover DV (1975) Characteristics of proteins in single and double endosperm mutants of maize. In: Bauman LF, Mertz ET, Caballo A, Sprague EW (eds) High Quality Protein Maize. Dowden, Hutchinson and Ross, Stroudsburg, pp 291-305
- Nelson OE, Mertz ET, Bates LS (1965) Second mutant gene affecting the amino acid pattern of maize endosperm proteins. *Science* 150:1469-1470
- Nevins JR (1983) The pathway of eukaryotic mRNA formation. *Annu Rev Biochem* 52:441-466
- Nute PE, Mahoney WC (1980) Complete primary structure of the β chain from the hemoglobin of a baboon, *Papio cynocephalus*. *Hemoglobin* 4:109-123
- Okayama H, Berg P (1982) High efficiency cloning of full-length cDNA. *Mol Cell Biol* 2:161-170
- Paulis JW, Wall JS (1971) Fractionation and properties of alkylated-reduced corn glutelin proteins. *Biochim Biophys Acta* 251:57-69
- Paulis JW, James C, Wall JS (1969) Comparison of glutelin proteins in normal and high-lysine corn endosperms. *J Agric Food Chem* 17:1301-1305
- Pedersen K, Bloom KS, Anderson JS, Glover DV, Larkins BA (1980) Analysis of the complexity and frequency of zein genes in the maize genome. *Biochemistry* 19:1644-1650
- Pedersen K, Argos P, Narayana SVL, Larkins BA (1986) Sequence analysis and characterization of a maize gene encoding a high-sulfur zein protein of M_r 15000. *J Biol Chem* 261:6279-6284
- Peterson GL (1977) A simplification of the protein assay method of Lowry et al. which is more generally applicable. *Anal Biochem* 83:546-556
- Phillips RL, McClure BA (1985) Elevated protein-bound methionine in seeds of a maize line resistant to lysine plus threonine. *Cereal Chem* 62:213-218
- Phillips RL, McClure BA, Wolf J, Gonschick W (1981) Seedling screening for L-phenylthiothreosulfonamide in maize. *Crop Sci* 21:601-607
- Rigny P, W. Diermann M, Rhodes C, Berg P (1977) Labeling deoxyribonucleic acid to high specific activity *in vitro* by microtransformation with DNA polymerase I. *J Mol Biol* 111:227-251
- Rigny P, Gamasz E, Viotto A, Soave C (1977) Heterogeneity of storage proteins in maize. *Plant* 109:115-123
- Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74:5463-5467
- Shure M, Wessler S, Fedoroff N (1983) Molecular identification and isolation of the *waxy* locus in maize. *Cell* 35:225-235
- Soave C, Rignetti PG, Lorenzoni A, Giannetta E, Salamini L (1976) Expression of the *opaque-2* gene at the level of zein molecular components. *Mol Biol* 21:67-75
- Soave C, Sunan N, Viotto A, Salamini L (1978) Linkage relationships between regulatory and structural gene loci involved in zein synthesis in maize. *Theor Appl Genet* 52:263-267
- Sodek L, Wilson CM (1971) Amino acid composition of protein isolated from normal *opaque-2* and *waxy-2* corn endosperms by a modified Osborne procedure. *J Agric Food Chem* 19:1142-1150
- Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* 98:503-532
- Viera J, Messing J (1987) Production of single-stranded plasmid DNA. *Methods Enzymol* 153:3-11
- Wahl GM, Stern M, Stark GR (1979) Efficient transfer of large DNA fragments from agarose gels to nitrobenzyl dimethyl-paper and rapid hybridization by using dextran sulfate. *Proc Natl Acad Sci USA* 76:3683-3687
- Wilson CM (1983) Seed protein fractions of maize, sorghum, and related cereals. In: Gonschick W, Nutter HP (eds) Seed Proteins: Biochemistry, Genetics, Nutritive Value. Martinus Nijhoff/Junk, The Hague, Netherlands, pp 273-307
- Wilson DR, Larkins BA (1984) Zein gene organization in maize and related grasses. *J Mol Evol* 20:336-340
- Wolf MJ, Khoo U, Seelinger HL (1967) Subcellular structure of endosperm protein in high-lysine and normal corn. *Science* 157:556-557
- Wood D (1984) Oligonucleotide screening of cDNA libraries. *Focus* 6:1-5
- Zimmerman CL, Appella E, Pisano JJ (1977) Rapid analysis of amino acid phenylthiohydantons by high-performance liquid chromatography. *Anal Biochem* 77:369-375

Communicated by G.R. Funk

Received July 26, 1987 / October 26, 1987

Cloning and sequence analysis of a cDNA encoding a Brazil nut protein exceptionally rich in methionine

Susan B. Altenbach, Karen W. Pearson, Filomena W. Leung & Samuel S. M. Sun
ARCO Plant Cell Research Institute, 6560 Trinity Court, Dublin, CA 94568, USA (for offprints)

Received 25 September 1986; accepted 28 October 1986

Keywords: Brazil nut, methionine-rich protein, oilseed proteins, seed storage proteins

Abstract

The primary amino acid sequence of an abundant methionine-rich seed protein found in Brazil nut (*Bertholletia excelsa* H.B.K.) has been elucidated by protein sequencing and from the nucleotide sequence of cDNA clones. The 9 kDa subunit of this protein was found to contain 77 amino acids of which 14 were methionine (18.0%) and 6 were cysteine (8.0%). Over half of the methionine residues in this subunit are clustered in two regions of the polypeptide where they are interspersed with arginine residues. In one of these regions, methionine residues account for 5 out of 6 amino acids and four of these methionine residues are contiguous. The sequence data verifies that the Brazil nut sulfur-rich protein is synthesized as a precursor polypeptide that is considerably larger than either of the two subunits of the mature protein. Three proteolytic processing steps by which the encoded polypeptide is sequentially trimmed to the 9 kDa and 3 kDa subunit polypeptides have been correlated with the sequence information. In addition, we have found that the sulfur-rich protein from Brazil nut is homologous in its amino acid sequence to small water-soluble proteins found in two other oilseeds, castor bean (*Ricinus communis*) and rapeseed (*Brassica napus*). When the amino acid sequences of these three proteins are aligned to maximize homology, the arrangement of cysteine residues is conserved. However, the two subunits of the Brazil nut protein contain over 19% methionine whereas the homologous proteins from castor bean and rapeseed contain only 2.1% and 2.6% methionine, respectively.

Introduction

In contrast to the seed proteins from many plants which contain relatively low levels of the sulfur-containing amino acids, the seed proteins from Brazil nut (*Bertholletia excelsa* H.B.K.) contain large percentages of methionine and cysteine, 8.3%–9.1% by weight [3, 26]. From a 2S albumin fraction of Brazil nut proteins, we previously purified an abundant sulfur-rich protein. This sulfur-rich protein consists of two low molecular weight subunits, a 9 kDa polypeptide and a 3 kDa polypeptide, which associate through disulfide bridges

to form a 12 kDa protein molecule (unpublished data). The sulfur-rich protein is synthesized in the seed only at a particular developmental stage, about 8 to 9 months after flowering. *In vitro* and *in vivo* labelling studies have indicated that this protein is synthesized initially as a larger precursor polypeptide of about 18 kDa which then undergoes three proteolytic processing steps before it attains its mature form [2].

We now report the amino acid sequence of some 77% of the large subunit of the sulfur-rich protein obtained by Edman degradation. Using a synthetic oligodeoxynucleotide probe whose sequence was

based on a methionine-rich region found in this partial amino acid sequence, we have identified cDNA clones encoding the sulfur-rich protein. In this paper, we present the complete nucleotide sequence of one Brazil nut cDNA clone and verify that the sulfur-rich protein encoded by this clone is synthesized as a larger precursor polypeptide. We have correlated the three processing steps by which the encoded polypeptide is sequentially trimmed to the 9 kDa and 3 kDa polypeptides with the sequence information and demonstrate that the 9 kDa subunit encoded by this clone contains 18% methionine and 8% cysteine. Finally, a computer search of available protein sequences revealed that the methionine-rich protein from Brazil nut is homologous in its amino acid sequence to small water-soluble seed proteins found in castor bean and rapeseed which contain only modest levels of methionine.

Materials and methods

Plant material

Brazil nuts are indigenous to the Amazon River basin; they do not grow anywhere in the United States. Brazil nut fruits were obtained approximately 9 months after flowering from Brazil (Manaus) or Peru (Iquitos or Puerto Maldonado).

Purification of the sulfur-rich protein and amino acid sequence determination

Brazil nut embryos were ground into a fine paste and defatted by extraction with hexane. The resulting defatted Brazil nut flour was then extracted in a buffer containing 1 M NaCl in 0.035 M sodium phosphate buffer, pH 7.5. The sulfur-rich protein was purified from this crude extract by the procedure of Youle and Huang [26]. The resulting sucrose gradient fractions were dialyzed extensively against deionized water at 4°C to precipitate the contaminating globulin proteins. The final protein sample contained polypeptides of 9 kDa and 3 kDa when analyzed on SDS-20% polyacrylamide gels.

The protein sample for sequencing was prepared by incubation of 2 mg of the purified sulfur-rich protein with 1 M Tris-HCl buffer, pH 8.5, containing 1 mM EDTA and 0.15 M 2-mercaptoethanol at 37°C under nitrogen gas for 4.5 hours. At the end of the incubation, iodoacetic acid was added to a final concentration of 0.22 M and the sample was incubated at 37°C in the dark for 30 minutes. After this treatment, the protein sample was dialyzed extensively against deionized water and lyophilized.

Sequence analysis of the sulfur-rich protein was performed by automated Edman degradation [6] on a Beckman 890C liquid-phase sequencer equipped with a cold trap using program 050783 with 0.1 M Quadrol (Beckman Instruments, Inc.) and polybrene (2 mg) as a carrier. About 10 nmol of the sulfur-rich protein were applied into the liquid phase sequencer. Norleucine was added to the fractions and used as an internal standard for quantitation of each cycle. Phenylthiohydantoin-amino acids were identified and measured by (Packard 419) gas liquid chromatography [20], (Water 6000A) high performance liquid chromatography [4] and thin layer chromatography [9]. At least two of these methods were used at each step. A total of 60 cycles of degradation were conducted, and 97% repetitive yield was observed. No PTH-amino acid could be identified after cycle 57.

Preparation of cDNA library and isolation of clones

Polyadenylated RNA was prepared from 8-month-old developing Brazil nut seeds by methods described previously for *Phaseolus vulgaris* [7] and was cloned in the linker-primer vector pARC7 [1]. The resulting clones were screened by colony hybridization [24] using a ³²P-labelled probe which consisted of a mixture of 6 synthetic oligodeoxynucleotides complementary to the 6 possible RNA sequences which could encode a methionine-rich region found in the partial amino acid sequence of the 9 kDa subunit of the sulfur-rich protein (Fig. 1B). The probe was hybridized to the filters in 6 × NET (0.9 M NaCl, 0.09 M Tris Cl, pH 7.5, 0.006 M EDTA), 0.05% NP-40, and 250 µg/ml yeast tRNA at 37°C for 20 hours. The filters were

washed in $6 \times$ SSC at 37°C before autoradiography.

Sequencing of cDNA and primer extension analysis

The sequence of cDNA clone pHS-3 was determined from both DNA strands by the dideoxy chain termination method [22]. Where necessary, regions of the clone were also sequenced by the method of Maxam and Gilbert [13]. The 25 nucleotides at the 5' end of the mRNA encoding the sulfur-rich protein were not represented in pHS-3 but were obtained by using a synthetic oligodeoxynucleotide complementary to nucleotides 449–669 as a primer to synthesize DNA complementary to the 5' end of the mRNA and sequencing the resulting extension product. For primer extension, the oligodeoxynucleotide 5'-AATCTTCGCCATGGT-GATTCT 3', labelled at its 5' end, was annealed to $3\ \mu\text{g}$ of poly(A)⁺ RNA from the seeds of 9-month-old Brazil nuts in 8 mM Tris pH 7.5, 5 mM EDTA at 90°C for 5 minutes. NaCl was added to 0.1 M and the sample was incubated for 20 minutes at 90°C followed by 15 minutes at 25°C . The annealed DNA sample was brought to a final concentration of 50 mM Tris pH 8.3, 5 mM DTT, 15 mM MgCl_2 , 0.5 mM dNTPs, and $0.1\ \mu\text{g/ml}$ BSA. AMV reverse transcriptase (BRL, 37.5 units) was added and the reaction was incubated at 37°C for 90 minutes. EDTA was added to 20 mM and the sample was extracted twice with phenol:chloroform:isoamyl alcohol (25:24:1) and precipitated with ethanol. After denaturation, the samples were subjected to electrophoresis on an 8% sequencing gel. Three bands resulted which differed in length by single nucleotides. DNA from each of the three bands was eluted from the gel and sequenced by the method of Maxam and Gilbert [13].

Hybrid-selected translation of cDNA clones

Characterization of cDNA clones by hybrid-selected translation was performed as described by Maniatis [12]. Three micrograms of either pHS-3 or

pAKC-7 plasmid DNA were denatured, bound to nitrocellulose paper and hybridized to $2\ \mu\text{g}$ of Poly(A)⁺ RNA prepared from 9-month-old Brazil nut seeds. RNA which was specifically bound to the DNA was then eluted, precipitated with ethanol along with $5\ \mu\text{g}$ carrier yeast tRNA, and translated in a wheat germ system [7]. In addition, the translation products directed by RNA selected by pHS-3 were immunoprecipitated [7] with a polyclonal antibody which had been made to a mixture of the 9 kDa and 3 kDa components of the mature Brazil nut sulfur-rich protein and its 12 kDa precursor. The proteins, labelled with [^{35}S]methionine, were analyzed on a SDS-20% polyacrylamide gel [11] and visualized by autoradiography.

Results

Partial amino acid sequence of the sulfur-rich protein

Two amino acid sequences were obtained from the analysis of the carboxymethylated sulfur-rich protein: one major (80%) and one minor (20%). The major sequence starts with Pro-Arg-Arg-Gly-Met... as NH_2 -terminal amino acids, while the minor one starts with Gly-Met... (Fig. 1A). The two protein sequences are identical in the region sequenced except that the minor one is three amino acids shorter than the major one at the NH_2 terminus; thus we were able to determine the first 57 amino acids for the major sequence and the first 54 amino acids for the minor one. The sulfur-rich protein consists of two subunits, a 9 kDa polypeptide and 3 kDa polypeptide. Both the 54 and the 57 amino acid sequences exceed the length of the 3 kDa polypeptide, thus these sequences must represent 9 kDa polypeptides, possibly two members of the 9 kDa polypeptide family. We did not obtain any amino acid sequence for the 3 kDa polypeptide, suggesting that either the 3 kDa polypeptide sequence is identical to the 9 kDa sequence or the NH_2 terminus of the 3 kDa polypeptide is blocked (see Discussion).

This amino acid sequence represents about 77% of the 9 kDa subunit. The sequence contains un-

A

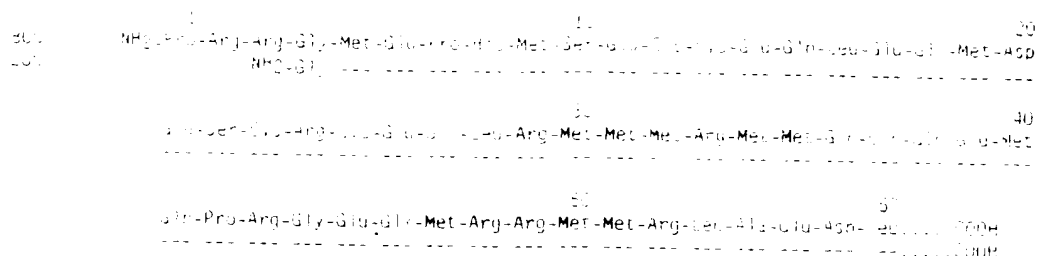


Fig. 1A. The partial amino acid sequence of the 9 kDa subunit of the sulfur-rich protein from Brazil nut. After reduction and carboxymethylation, the purified sulfur-rich protein was sequenced using an automatic liquid-phase sequencer. Two sequences, one major (80%) with Pro as the NH₂ terminal amino acid (shown in the first line), and one minor (20%) with Glu as the NH₂ terminal amino acid (shown in the second line), were detected. Amino acid residues found in the minor sequence which are identical to those found in the major sequence are indicated with dashes. Methionine-rich regions found in the partial sequence are highlighted.

B

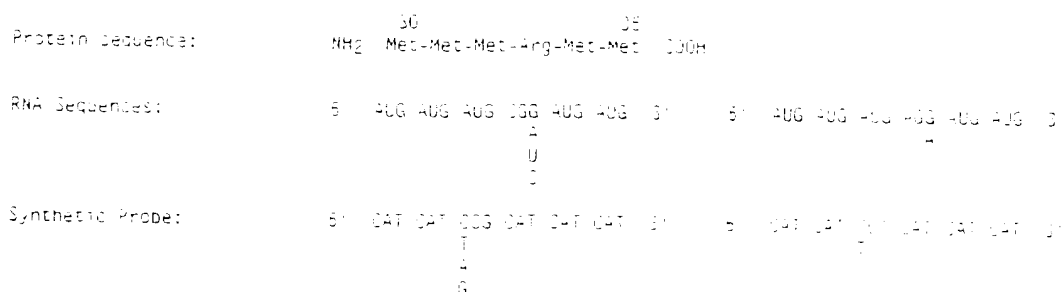


Fig. 1B. Amino acid sequence of the first methionine-rich region which was used as a basis for a synthetic oligodeoxynucleotide probe. The sequences of the 6 possible mRNAs encoding this portion of the protein sequence are shown in the second line and the sequences included in the synthetic oligodeoxynucleotide probe complementary to the mRNA are shown in the bottom line.

usually high levels of the sulfur amino acids: 21% methionine and 7% cysteine. There are two regions in the partial amino acid sequence where methionine residues are clustered with arginine residues: residues #29–35 (Arg–Met–Met–Met–Arg–Met–Met) and residues #47–52 (Met–Arg–Arg–Met–Met–Arg) (Fig. 1A).

Identification and characterization of cDNA clones encoding the sulfur-rich protein

An oligodeoxynucleotide probe was synthesized (by Biosearch, Inc.) which was complementary to the 6 possible RNA sequences encoding one of these methionine-rich regions (amino acid residues

#30–35) (Fig. 1B). This oligodeoxynucleotide probe hybridized to a number of clones from a cDNA library prepared using RNA from 9-month-old Brazil nut seeds. Twelve of these clones with inserts ranging from 350 bp to 700 bp were selected for further analysis.

Sequence analysis of one of these clones, pHS-3, demonstrates unequivocally that this cDNA encodes a polypeptide which is extremely rich in the sulfur-containing amino acids (Fig. 2A). The sequence of pHS-3 is 599 nucleotides long excluding the poly(A) tail. By primer extension analysis using a 21 base synthetic oligodeoxynucleotide complementary to a region near the 5' end of pHS-3, we determined that this cDNA clone falls 25 nucleotides short of the 5' end of the mRNA en-

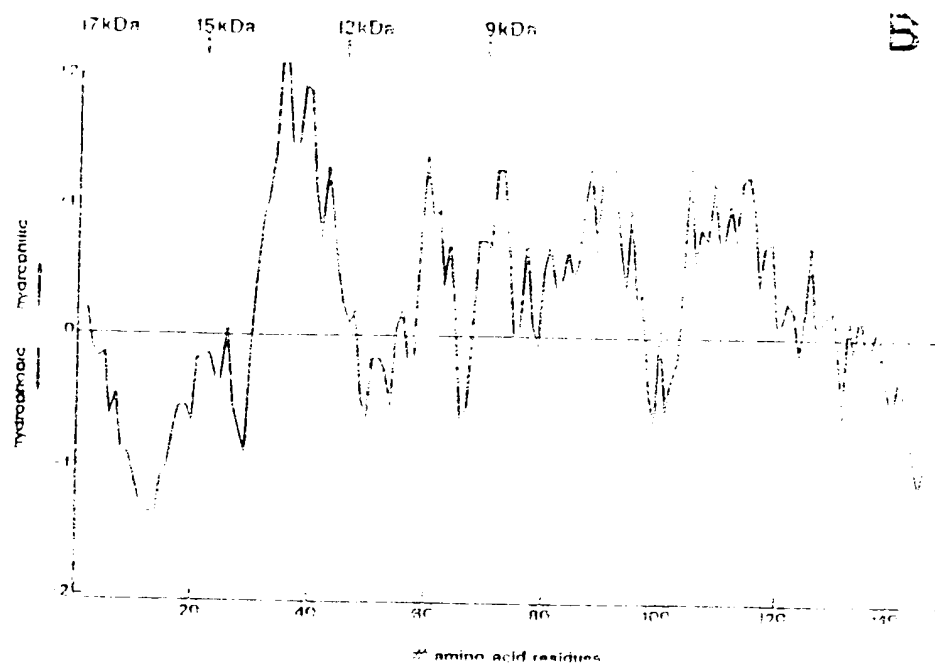


Fig. 2B. Hydropathy plot of Brazil nut sulfur-rich protein encoded by cDNA clone pHS-3. Plot showing the hydrophobic and hydrophilic regions of the protein encoded by pHS-3 was generated using the algorithm of Hopp and Wood [9]. The horizontal line in the middle of the plot represents a hydrophilicity value of 0. Hydrophilic regions are plotted above the 0-line and hydrophobic regions beneath the 0-line. Numbers along the x-axis refer to the number of amino acid residues from the NH_2 terminus of the protein. The approximate location of cleavages believed to be involved in the maturation of the Brazil nut sulfur-rich protein are shown with arrows.

coding the sulfur-rich protein. There are no ATG codons in the sequence of the primer extension product; thus, the first ATG codon found in pHS-3 (residues #58–60) represents the initiation codon for protein synthesis. This ATG fits with the consensus sequence for eucaryotic protein initiation sites [10]. A stop codon, TGA, is encoded by nucleotides #495–497 in pHS-3. The resulting open reading frame could encode a polypeptide of 146 amino acids, of which over 20% are sulfur-containing amino acids; 15.1% of these residues are methionine while 5.5% are cysteine. The first portion of the encoded polypeptide contains a large proportion of hydrophobic residues: of the 22 residues at the amino terminus of the protein, 36% are alanine and 18% are leucine. In comparison, the rest of the polypeptide is rich in arginine, glutamine and glutamic acid, a composition which is characteristic of other plant seed storage proteins. A hydropathy plot (Fig. 2B) demonstrates that the amino terminus of the polypeptide is hydrophobic

while the remainder of the polypeptide is largely hydrophilic.

By aligning the amino acid sequence derived from the nucleotide sequence with the major sequence determined from the purified 9 kDa subunit, we have found that the coding region for the 9 kDa polypeptide begins 265 nucleotides from the 5' end of the mRNA. By adding up the molecular weights of the individual amino acids encoded by this region, we arrive at a value of almost 9 kDa. The amino acid sequence derived from the nucleotide sequence of the portion of the open reading frame between nucleotides 265 and 425 agrees quite well, although not precisely, with the major 57 residue partial amino acid sequence of the 9 kDa subunit of the sulfur-rich protein (Fig. 2A). Methionine residues are very predominant in the 9 kDa subunit of the mature protein. There are 14 methionine residues in this region, representing 18.2% of the 77 amino acid polypeptide. Eight of these 14 methionines are found clustered with

arginine residues in two regions of the polypeptide. In the first cluster, between amino acid residues #99 and 104, five out of six residues are methionines and four of the methionine residues are contiguous. The second methionine cluster, between amino acid residues #116 and 121, includes three methionine residues and three arginine residues. Interestingly, 2 of the 4 amino acid differences which are found between the amino acid sequence determined from the protein and that derived from the nucleotide sequence are found in the methionine-rich region that was used as a basis for the synthetic oligodeoxynucleotide probe. A second cDNA clone selected by the same probe was perfectly homologous with one of the sequences represented in the probe (unpublished data), suggesting that the sulfur-rich protein is encoded by a family of genes with some variation in these methionine-rich regions. The 9 kDa subunit of the Brazil nut protein also contains a high proportion of cysteine (7.7%).

By hybrid-selected translation, we have found that pHS-3 is able to select a mRNA from a population of 9-month-old Brazil nut RNAs which directs the synthesis of an 18 kDa polypeptide *in vitro* (Fig. 3). This 18 kDa polypeptide is immunoprecipitable with a polyclonal antibody raised in rabbits against the purified Brazil nut sulfur-rich protein, demonstrating conclusively that the sulfur-rich protein is synthesized initially as a larger precursor polypeptide.

Homology of the sulfur-rich protein to other water-soluble seed proteins

In a computer search of proteins whose amino acid sequences have been determined, we found that the sulfur-rich protein from Brazil nut shares a great deal of homology with both the large and the small subunits of a low molecular weight and water-soluble seed storage protein from castor bean (*Ricinus communis*) [23]. We have aligned the amino acid sequence of the small subunit of this castor bean protein with the Brazil nut sequence starting at amino acid residue #35 and that of the large subunit of the castor bean protein with amino acid

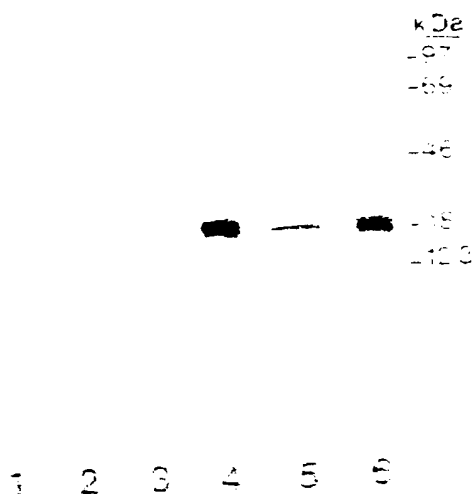


Fig. 3. Identification of a cDNA clone for the sulfur-rich Brazil nut protein by translation of hybrid-selected mRNAs. Lane 1 shows endogenous proteins synthesized in the wheat germ system and labelled with [35 S] methionine. Lanes 2 and 3 show labelled proteins synthesized by either RNA selected by the vector pARCT or 5 µg yeast tRNA. The translation products of RNA selected by pHS-3 are displayed in lane 4 and these products are immunoprecipitated with the Brazil nut sulfur-rich protein antibody in lane 5. Lane 6 shows the total translation products of Brazil nut poly(A)⁺ RNA in the wheat germ system.

residues starting at #72 (Fig. 4A). Allowing 2 small gaps in the small subunit comparison and 4 small gaps in the large subunit comparison to maximize sequence homology, we find over 44% homology between the castor bean protein and the Brazil nut sulfur-rich protein. Both proteins are high in glutamine, glutamic acid and arginine (22% and 13% for the Brazil nut protein and 29.5% and 10.5% for the castor bean protein, respectively), and the positions of many of these residues are conserved in the two proteins. Interestingly, both the Brazil nut and the castor bean proteins are relatively rich in cysteine (7% and 8.4%, respectively) and the positions of these residues are similar in both proteins. Another small water-soluble protein found in rapeseed (*Brassica napus*), napin [5], shows some homology (about 21%) with the Brazil nut protein (Fig. 4A).

A

Brazil Nut
Castor Bean
Rapeseed

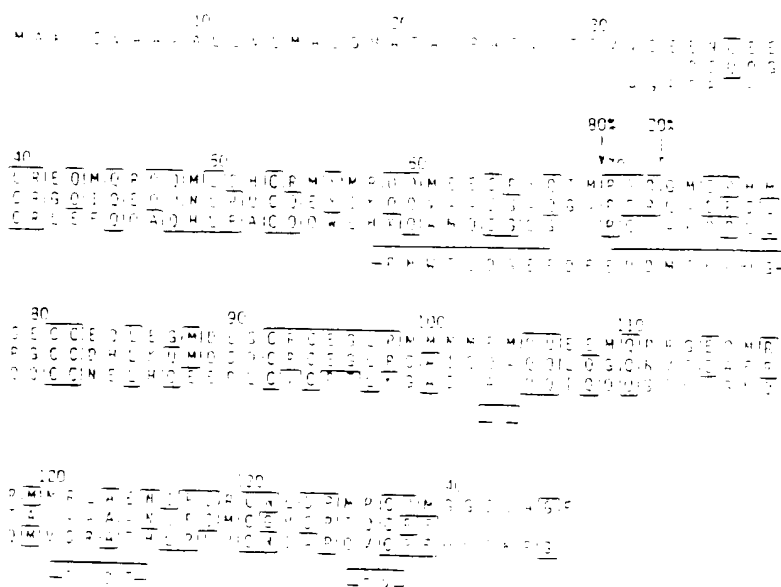


Fig. 4A. Comparison of the amino acid sequences of the 2S sulfur-rich protein of Brazil nut and the 2S water-soluble seed protein from castor bean and rapeseed. The top line shows the amino acid sequence of the precursor polypeptide for the Brazil nut sulfur-rich protein (derived from the nucleotide sequence of cDNA clone pHS-3). The amino acid sequences of both subunits of the castor bean protein [23] are shown in the second line and the amino acid sequences of both subunits of napin (derived from the sequences of napin cDNA clones) [5] are shown in the third line. The one letter amino acid code was used for these comparisons and the sequences were aligned to maximize homology with the Brazil nut protein. Homologous residues are enclosed in boxes. The locations of the first processing sites involved in the maturation of the Brazil nut sulfur-rich protein (major, 80%, and minor, 20%) were determined from the amino acid sequence analysis of the 9 kDa subunit and are indicated with arrows.

Although the homology noted between Brazil nut and rapeseed proteins is substantially less than that between the Brazil nut and castor bean proteins, 24 out of 28 amino acids (85.7%) conserved between the Brazil nut and rapeseed proteins are also common to the castor bean protein. In addition, the positions of cysteine residues in all three proteins are conserved. However, the Brazil nut protein is unusually rich in methionine (19%) while the castor bean and rapeseed proteins contain only about 2% methionine. Thus, a large percentage of the non-homology between the Brazil nut protein and the castor bean or rapeseed protein sequences is due to differences in their methionine contents. We have also compared the protein sequence of the Brazil nut sulfur-rich protein to that of the 15 kDa high sulfur zein protein from maize [18] which contains about 11% methionine and have found no significant homology between these two proteins.

Discussion

The majority of known proteins, of both plant and animal origin, have relatively low levels of methionine, usually around 1–2% as predicted by the theory of molecular evolution [16]. In the present study, we have partially sequenced an abundant protein from Brazil nuts which is exceptionally rich in methionine (18%) and have identified and sequenced a cDNA clone encoding this protein. Only one plant protein with comparable levels of methionine has been reported in the literature. Phillips and McClure recently described the amino acid composition of a polypeptide of 10 kDa in a maize mutant, BSSS-53, containing 21 mol% methionine [19].

The sequence data from the Brazil nut cDNA clone as well as the data from the hybrid-selected translation experiment are consistent with previous

in vitro translation studies which have shown that the sulfur-rich protein is synthesized as a larger precursor polypeptide. The size of the polypeptide encoded by pHS-3 would be about 17 kDa, which is close to the 18 kDa value for the precursor obtained from the sizing on polyacrylamide gels of the polypeptides translated from Brazil nut RNA *in vitro* [2]. The correlation of the amino acid sequence obtained from the purified 9 kDa subunit with the last 77 amino acids of the sequence derived from the nucleotide sequence indicates that the processing steps which are involved in the maturation of the sulfur-rich protein must be taking place at the amino terminal end of the precursor. Previous *in vivo* labelling studies demonstrated that there are 3 distinct processing steps. First, a small peptide, most likely a signal sequence, is cleaved from the 18 kDa precursor to generate a 15 kDa polypeptide which subsequently is processed to a 12 kDa polypeptide and then to the 9 kDa and 3 kDa subunits [2]. We have not determined experimentally the precise residues which are cleaved upon maturation of the sulfur-rich protein. Nonetheless, we can propose approximate cleavage sites that would divide the amino acid sequence into four domains corresponding to the observed polypeptides (Fig. 2A). The hydrophobic nature of the amino terminus of the encoded polypeptide (Fig. 2B) suggests that this region serves as a signal peptide. The alanine and phenylalanine residues at positions #22 and 23 would represent a possible cleavage site for a signal peptidase as determined by the $(-3-1)$ rule of Von Heijne [25]. A second cleavage may take place around amino acid residue #46 and would result in a polypeptide of about 12 kDa. We have attempted to determine the exact location of this cleavage site by sequencing the 12 kDa precursor polypeptide, but found that its NH_2 terminus is blocked. Finally, the major (80%) partial amino acid sequence of the 9 kDa subunit would predict that the cleavage site for the third processing step is between methionine residue #69 and proline residue #70, whereas the minor sequence (20%) would indicate that the final processing site is three amino acids away, between residues #72 and 73. The 3 kDa region clipped off in this final processing step is extremely rich in methionine

($\sim 20\%$). We suspect that this portion of the precursor accumulates and gives rise to the 3 kDa subunit of the sulfur-rich protein. *In vivo* labelling studies using ^{35}S methionine indicate that the 3 kDa subunit is rich in methionine (data not shown). In addition, the amino acid composition of the sulfur-rich protein supports this notion. Tyrosine and threonine residues are present in the amino acid analysis of the purified sulfur-rich protein (9 kDa + 3 kDa) (data not shown). These residues are not found in the amino acid sequence derived from the nucleotide sequence of the 9 kDa subunit but are present in the 3 kDa region immediately preceding the 9 kDa subunit.

The homology between the sulfur-rich protein from Brazil nut and seed proteins from castor bean and rapeseed is particularly striking since the three plants are not closely related taxonomically and the castor bean and rapeseed proteins contain low levels of methionine. The proteins from all three plants consist of a small and a large subunit polypeptide and contain high levels of cysteine. The positions of these cysteine residues are conserved, suggesting that the structural frameworks of these three proteins may be quite similar despite the dramatic differences in their methionine contents. This structural similarity may be conserved in the small water-soluble proteins in other oilseeds of diverse phylogenetic relationships as well. In a survey of the amino acid compositions of 25 seed proteins, Youle and Huang [26] noted that the levels of cysteine in proteins from different oilseeds (sunflower, mustard, linseed, lupin, cucumber, Brazil nut, hazelnut, yucca, castor bean, and cotton) were quite high and in fact very similar. Because of their high amide contents, abundance in seeds, and disappearance from seeds during germination, these low molecular weight proteins were suggested to function as seed storage proteins with the additional and unique role of providing sulfur reserves for germination [26]. Of these proteins, however, only the Brazil nut 2S protein contains unusually high levels of methionine, contrary to theoretical predictions based on the theory of molecular evolution [16]. At the present time, we do not know why Brazil nuts might require such high levels of methionine. The soil in the Amazon region is rather poor

in sulfur [21]; possibly these levels of methionine are required in order to provide an adequate supply of methionine to the germinating seeds. Whatever the function of the Brazil nut sulfur-rich protein, it appears that the structural framework of the 2S seed proteins is flexible enough to accommodate large numbers of methionine residues while still preserving the small size, water solubility, and high amide content of these proteins.

Both the castor bean protein and the rapeseed protein are analogous to the Brazil nut sulfur-rich seed protein in that they are composed of two low molecular weight subunits. In the case of castor bean, the large subunit of the protein is homologous with the 9 kDa subunit portion of the Brazil nut precursor polypeptide while the small subunit of the castor bean protein appears to correspond to the region of the Brazil nut protein which we believe encodes the 3 kDa subunit. Interestingly, the junction between the large and small subunits of the castor bean protein corresponds to the minor cleavage site of the Brazil nut 12 kDa precursor (amino acid residue #72) (Fig. 4B). These data suggest that both subunits of the castor bean protein

may be synthesized as part of a larger precursor, similar to the Brazil nut sulfur-rich protein and that the final processing step involved in the maturation of the castor bean protein may be similar to that found with the Brazil nut protein.

The processing involved in the maturation of the rapeseed protein [5] also bears similarities to that of the Brazil nut sulfur-rich protein. As with the Brazil nut protein, the large subunit of napin is found at the carboxyl terminal portion of the precursor (Fig. 4B). In both Brazil nut and rapeseed, the precursor polypeptide undergoes extensive processing before reaching its mature subunits. From the best alignment of the amino acid sequence of the large subunit of the rapeseed protein with that of the Brazil nut protein sequence, it appears that the cleavage site of the large subunit of napin occurs at about the same point as the primary cleavage site of the Brazil nut large subunit (amino acid residue #69) (Fig. 4B).

In the past, there has been much effort to enhance the sulfur amino acid content of seeds, particularly those from legumes, by conventional plant breeding approaches. The overall improvement in

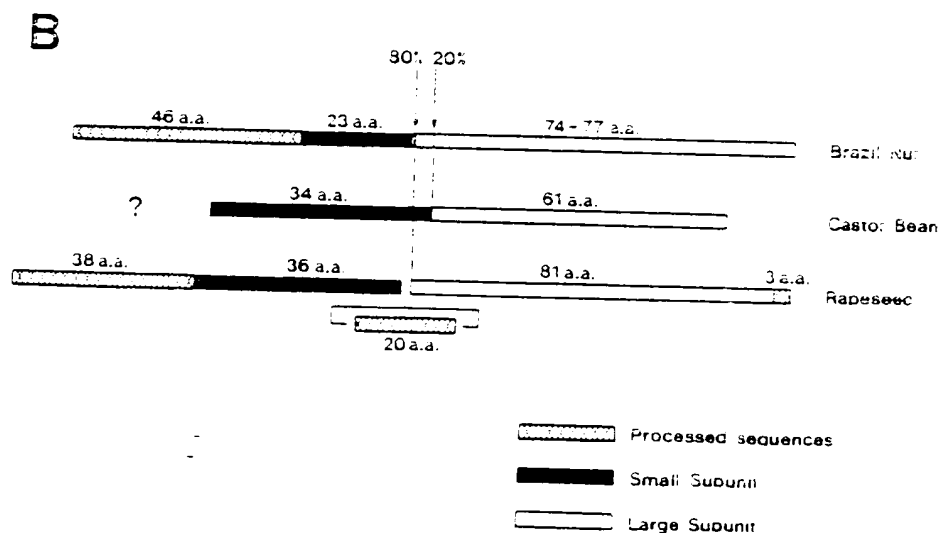


Fig. 4B. Comparison of the processing schemes utilized in the maturation of the subunits of the 2S water-soluble seed proteins of Brazil nut, castor bean and rapeseed showing the positions of the large and small subunits of the castor bean and rapeseed proteins relative to the Brazil nut sulfur-rich protein. The locations of the final processing sites involved in the maturation of the Brazil nut sulfur-rich protein are indicated with arrows. The large subunit of the castor bean protein begins at the same amino acid residue as 20% of the large subunit molecules from the Brazil nut protein, whereas the processing site in the rapeseed protein appears to correlate with the 80% processing site from Brazil nut.

the nutritional quality of these seeds has not been significant [17], although the same approach was successful in obtaining high lysine corn [14, 15]. Studies of seed proteins in oilseeds have shown that there is a wide occurrence of abundant 2S proteins in diverse plant species. These proteins appear to have similarity in their structural framework and precursor processing, seem to serve a storage function, and have a seemingly flexible amino acid composition. The fact that a large amount of methionine is localized in a single 2S protein species in Brazil nut suggests to us a molecular approach for improving the nutritional quality of seed proteins deficient in the sulfur amino acids. The cloning of a cDNA encoding this sulfur-rich protein thus represents a first step in an effort to alter the amino acid composition of seed proteins. A further understanding of the genes which encode this unusual sulfur-rich protein should provide additional useful information.

Acknowledgements

We thank Mr. Bruce Nelson for help in collecting the Brazil nut fruits, Mr. Alan Smith (Department of Biochemistry, University of California, Davis, CA) for protein sequence determination, Ms. Kathy Mead and Dr. Danny Alexander for the synthesis of the oligodeoxynucleotide used in the primer extension analysis, and Dr. Phil Filner for helpful suggestions and critical review of this manuscript.

References

1. Alexander DC, McKnight TD, Williams BG. A simplified and efficient vector-primer cDNA cloning system. *Gene* 31:79-89, 1984.
2. Altenbach SB, Pearson RW, Leung FW, Sun SS-M. The step-wise processing of a sulfur-rich seed protein from Brazil nut (*Bernhartia excelsa*). In: Shannon LM, Chrispeels MJ (eds) *Molecular Biology of Seed Storage Proteins and Lectins*. Waverly Press, Baltimore, Maryland, 1986, pp 137-146.
3. Antunes AJ, Markakis PJ. Protein supplementation of navy beans with Brazil nuts. *Agric Food Chem* 25:1096-1098, 1977.
4. Brown AS, Moie JE, Weissinger A, Bennett JC. Methanol solvent system for rapid analysis of phenylthiohydantoin amino-acids by high pressure liquid chromatography. *J Chromatogr* 14:122-132, 1975.
5. Crouch ME, Tenberge KM, Simon AE, Per R. cDNA clones for *Bernhartia nut* seed storage protein: evidence from nucleotide sequence analysis that both subunits of napin are cleaved from a precursor polypeptide. *Mol & Appl Genetic* 2:273-282, 1982.
6. Edman P, Begg G. A protein sequencer. *Eur J Biochem* 150:9-20, 1967.
7. Hall TC, Xia Y, Buchbinder BL, Pine JW, Sun SS-M, Bhatnagar M. Messenger RNAs for 2S protein in Brazil nut seed: cell-free translation and product characterization. *Proc Natl Acad Sci USA* 75:2196-2200, 1978.
8. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci USA* 76:3624-3628, 1979.
9. Kulb, DK. Micro-boramide thin layer chromatography of phenylthiohydantoin amino-acids as subunit amino acids: a rapid micro-technique for simultaneous multi-sample identification after automated Edman degradation. *Anal Biochem* 56:564-572, 1974.
10. Kozak M. Comparison of initiation of protein synthesis in prokaryotes, eukaryotes, and organelles. *Microbiol Rev* 47:1-45, 1981.
11. Laemmli UK. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* 227:680-685, 1970.
12. Maniatis T, Fritsch EF, Sambrook J. *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory, New York, 1982, pp 331-332.
13. Maxam AM, Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci USA* 74:560-564, 1977.
14. Merz ET, Bates LS, Nelson OE. Mutant gene that changes protein composition and increases lysine content of maize endosperm. *Science* 147:279-280, 1964.
15. Nelson OE, Merz ET, Bates LS. Second mutant gene affecting the amino acid pattern of maize endosperm proteins. *Science* 150:1469-1470, 1965.
16. Ohta T, Kimura M. Amino-acid composition of proteins as a product of molecular evolution. *Science* 174:150-153, 1971.
17. Payne PJ. Breeding for good in quantity and in quality of seed crops. In: Daussant J, Messe J, Vaughan J (eds) *Seed Protein*. Academic Press, Inc., London, 1982, pp 223-253.
18. Pedersen K, Argos P, Narayana SVL, Larkins B. Sequence analysis and characterization of a maize gene encoding a high-sulfur seed protein. *Proc Natl Acad Sci USA* 83:6279-6284, 1986.
19. Phillips L, McClure BA. Elevated protein-bound methionine in seeds of a maize *Zea mays* line resistant to lysine plus threonine. *Cereal Chem* 62:213-218, 1985.
20. Pisano JJ, Brontzer TJ, Brewer HB Jr. Advances in the gas chromatographic analysis of amino-acid phenylthiohydantoin and methylthiohydantoin. *Anal Biochem* 45:43-59, 1972.

21. Sanchez PA, Bandy DE, Villachica JH, Nicholas JJ: Amazon basin: management for continuous crop production. *Science* 216:821-827, 1982.
22. Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74:5463-5467, 1977.
23. Sharief FS, Li SS-L: Amino acid sequence of small and large subunits of seed storage protein from *Ricinus communis*. *J Biol Chem* 257:14753-14759, 1982.
24. Taub E, Thompson EB: An improved method for preparing large arrays of bacterial colonies containing plasmids + hybridization: *in situ* purification and stable bonding DNA on paper filters. *Anal Biochem* 126:222-230, 1982.
25. Von Heine G: How signal sequences maintain cleavage specificity. *J Mol Biol* 173:243-251, 1984.
26. Youle RJ, Huang AH: Occurrence of low molecular weight and high cysteine-containing albumin storage proteins in oilseeds of diverse species. *Amer J Bot* 68:44-48, 1981.

Volume 17 Number 5

The nucleotide sequence

H. Luerssen, W. M. Ma.

Institut für Humangenetik

EMBL accession no. X1458

Institut für Humangenetik
submitted April 4, 1989

We screened a 81 mer prepared sequence of the 3 independent c containing the protein 2. Rat The insert of t region of 155 b position 468. T aminoacids of r deduced aminoac There is a homo polypeptides of aminoacid seque aminoacids in p

CGATGAGTGGTGGGAAGGCTCTGC:

70 80

CGG CCT CAA AGT CAC ACC A
Arg Pro Gln Ser His Thr

160 170

CCC AGC CCT GGC CCG CCG
Pro Ser Pro Gly Pro Pro

250 260

AAG AAC AGG AAG ACC TTG
 CYS ASN ARG LYS THR LEU

340 351

GGA UGA AGA TAC AAG TGA
Gly Arg Arg Tyr Lys ***
116

450 460 470

SCIENTIFIC PRACTICES

GCATTCTATGCAACATGGATTAA
Acknowledgement:

References

References

1. Kenneth D. Co

Biochem. Biophys.

2. Kenneth C. K.

J. Biol. Chem.

6. _____

CTRL Press

© 2005 Blackwell Publishing Ltd, *Journal of Internal Medicine* 258: 105–112

3584

© IRL Press

Sequence and expression of a gene encoding an albumin storage protein in sunflower

R.D. Allen*, E.A. Cohen**, R.A. Vonder Haar, C.A. Adams, D.P. Ma, C.L. Nessler, and J.L. Thomas

Biology Department, Texas A & M University, College Station, TX 77843, USA

Summary. The complete sequence of a sunflower (*Helianthus annuus*) gene, *HuG5*, encoding a 2 S albumin storage protein was determined. The predicted unprocessed precursor has 295 amino acids, is rich in glutamine residues (24%) and contains a hydrophobic amino-terminus that is similar to the consensus signal peptide. Amino acid sequencing of the mature protein revealed extensive post-translational processing. Nuclease protection and primer-extension analysis indicated a major transcriptional start 50 nucleotides 5' of the predicted ATG start codon. Additional sequence data, determined from a nearly full length cDNA recombinant, indicate that *HuG5* is a member of a small gene family comprised of at least two divergent genes. Comparison of the predicted *HuG5* gene product with sequences of other known plant proteins revealed distant but significant homology with the napins of *Brassica* and other heterogeneous seed proteins in the albumin superfamily.

Key words: Sunflower · Albumin gene · DNA sequence

Introduction

The structure and expression of plant storage protein genes have been investigated in a number of monocot and dicot plant species (reviewed in Kreis et al. 1985a; Casey et al. 1986). In all cases, the accumulation of storage proteins during seed development and maturation requires the highly regulated expression of genes encoding these proteins and as such provides an excellent opportunity for analysis of the molecular mechanisms controlling ontogenic gene expression in plants. Sunflowers are particularly useful for these studies because the central disk of the sunflower inflorescence consists of hundreds of individual flowers, each of which produces a single embryo; consequently, a single sunflower plant can yield gram quantities of developmentally staged embryos.

Sunflower embryos accumulate two major classes of storage proteins. These are the 11 S globulins, soluble in 1 M NaCl, and 2 S albumins, soluble in water (Youle and

Huang 1981). The sunflower 11 S storage protein, designated heliantin (Schwenke et al. 1979), is structurally similar to legumin-like seed proteins of other plant species and is represented in plants by an approximately 500 kDa hexameric holoprotein. Each subunit of the holoprotein consists of a larger α polypeptide (30–40 kDa) and a smaller β polypeptide (23–27 kDa) linked by disulfide bonds (Gallarrondo et al. 1984); the heliantin α and β subunits are generated proteolytically from a larger precursor polypeptide (Higgins 1984). The cloning and expression of heliantin mRNAs have been described (Allen et al. 1985).

The synthesis, processing and accumulation of 2 S albumin seed proteins have been studied intensively in *Brassica napus* (Crouch et al. 1983; Ericson et al. 1986), pea (Higgin et al. 1986), radish (Laroche-Fajard and Delsens 1985), castor bean (Ford 1985) and Brazil nut (Sun et al. 1987). A major conclusion of these studies is that the characteristic low molecular weight, disulfide-linked albumin polypeptides found in mature seeds result from the extensive processing of larger precursors synthesized during embryogenesis. Two additional characteristics that define the 2 S albumin seed storage proteins are high amide content and high frequency of cysteine residues (Youle and Huang 1981).

In sunflower, the 2 S albumins represent more than 50% of the protein present in seeds (Youle and Huang 1981) and consist of two or three closely related polypeptides with molecular weights of approximately 19 kDa (Cohen 1986; Allen et al. 1987). The sunflower albumins apparently maintain a compact structure by intramolecular disulfide bonds resulting in a rapidly migrating species with an apparent molecular weight of 14 kDa when analyzed by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) under non-reducing conditions. When reduced, this species migrates as a 19 kDa polypeptide (Cohen 1986). In contrast, most other 2 S proteins are composed of large and small subunit polypeptides, derived from a single precursor, and linked by intermolecular disulfide bonds (Crouch et al. 1983; Ericson et al. 1986; Sun et al. 1987).

Albumin polypeptides can be detected in sunflower embryos by 5 days post-fertilization (DPE), 2 days before heliantin is detectable, and continue to accumulate through seed maturation. Sunflower albumin mRNAs, also first detected at 5 DPE, accumulate rapidly in sunflower embryos reaching maximum prevalence between 12 and 15 DPE. After this time albumin transcripts decrease in prevalence with kinetics similar to that observed for heliantin mRNA (Allen et al. 1987). Functional sunflower al-

* Present address: Department of Biology, Washington University, St. Louis, MO 63130, USA

** Present address: Department of Genetics, University of Georgia, Athens, GA 30602, USA

Offprint request to: J. L. Thomas

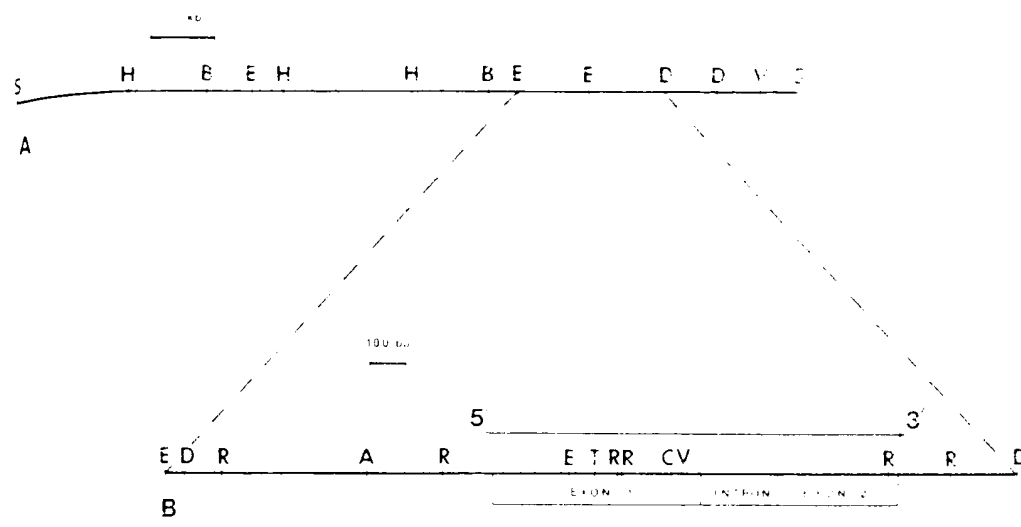


Fig. 1A, B. Physical map of sunflower albumin gene, *HaG5*. A Partial restriction map of a 1.5 kb sunflower genomic region including the *HaG5* transcription unit. B Detailed restriction map of the 2.2 kb region sequenced, indicating location of *HaG5* transcription unit. The solid bar beneath the map indicates the location of the 330 bp *EcoRI*/*RsaI* probe used for nuclease protection experiments. Restriction sites: A, *Acl*; B, *Bgl*II; C, *Cla*I; D, *Dra*I; E, *EcoRI*; H, *Hind*III; R, *Rsa*I; S, *Sac*I; T, *Sst*I; V, *Xba*I.

The nucleotide sequence of *HaG5* and the predicted amino acid sequence are shown in Fig. 2. An open reading frame (ORF) encoding a putative albumin storage protein begins with an ATG at position 888; this ORF continues for 575 nucleotides where it is interrupted by a 190 nucleotide intervening sequence. Placement of the intron was based on the discontinuity of the ORF in this region, on the presence of excellent consensus splice junctions and, most importantly, on the colinearity of the *HaG5* and *Ha5* sequences on either side of the intervening sequence. This single intron splits the AGG codon for arginine at amino acid position 192. Following the intron, the ORF continues for an additional 310 nucleotides where it is terminated by a TGA stop codon at position 1964. A consensus polyadenylation signal, AATAAA, is located 23 nucleotides 3' of the stop codon at position 1990. The combined length of exons 1 and 2 is 885 nucleotides indicating a protein coding capacity of 295 amino acid residues.

Transcript mapping

The site of transcriptional initiation for the *HaG5* transcription unit was determined by nuclease protection and by primer extension analysis. For nuclease protection, a 5' end-labeled, 330 bp *RsaI*/*EcoRI* fragment (position 758 to 1087 in Fig. 2; also see Fig. 1B) was hybridized with sunflower embryo or leaf RNA and subsequently digested with either S1 or mung bean nuclease, resulting nuclease-resistant fragments were resolved on sequencing gels. Results from one such experiment (Fig. 3A) revealed a major mung bean nuclease resistant DNA fragment 230 nucleotides (nt) in length. The same sized fragment is resistant to S1 nuclease (Fig. 3B). A second putative nuclease-resistant fragment is also observed in Fig. 3A at 298 nt. It is unlikely, however, that this fragment defines the 5' boundary of the *HaG5* transcription unit since the 298 nt molecular species is not observed when S1 nuclease is substituted for mung bean nuclease (see Fig. 3B). There are no detectable nuclease-resistant fragments generated when sunflower leaf RNA is hybridized with the 330 nucleotide *HaG5* probe. Primer extension analysis (data not shown) is consistent with a

transcription start site defined by the 230 nt nuclease-resistant DNA fragment. Taken together, these results suggest the transcriptional start site is located at position 858 (see Fig. 2).

Predicted protein characteristics

The calculated molecular weight for the unprocessed *HaG5* gene product is 38 kDa. The amino-terminal 27 residues are highly hydrophobic with an average hydrophobicity of -0.845 . It is likely but unproven that this hydrophobic domain is a signal sequence which facilitates transport of this protein into protein bodies. This "leader" sequence is probably removed during subsequent post-translational events. Using the rules defined by von Heijne (1986), we predict that the most likely site for cleavage of this putative signal sequence is after the alanine at residue 20 (see arrow Fig. 4).

Protein sequencing confirmed that *HaG5* encodes a major sunflower albumin storage protein and further demonstrated that the mature protein is the result of substantial post-translational processing. The major sunflower albumin was partially purified from mature seeds by chromatography on DEAE-cellulose. Because of its high pI, albumin was the only major seed protein that failed to bind to the column. Twenty micrograms of the unbound protein was further resolved on 10% SDS-PAGE and transferred to an aminopropyl-derivatized glass filter (Aehersold et al. 1986). The sequence of the first 12 residues beginning at the mature N-terminus was determined in the Texas A&M Biotechnology Support Laboratory. This sequence, indicated by the box in Fig. 4, is a perfect match with the amino acid sequence predicted from *HaG5* and would be expected to occur on a random basis at a frequency less than 10^{-12} .

The predicted amino acid composition of the mature sunflower albumin is compared with that of its precursor in Table 1. As expected from the amino acid composition reported for sunflower 2S albumins (Youle and Huang 1981), the mature sunflower protein is very glutamine rich (25%) and also has relatively high levels of cysteine (6.7%).

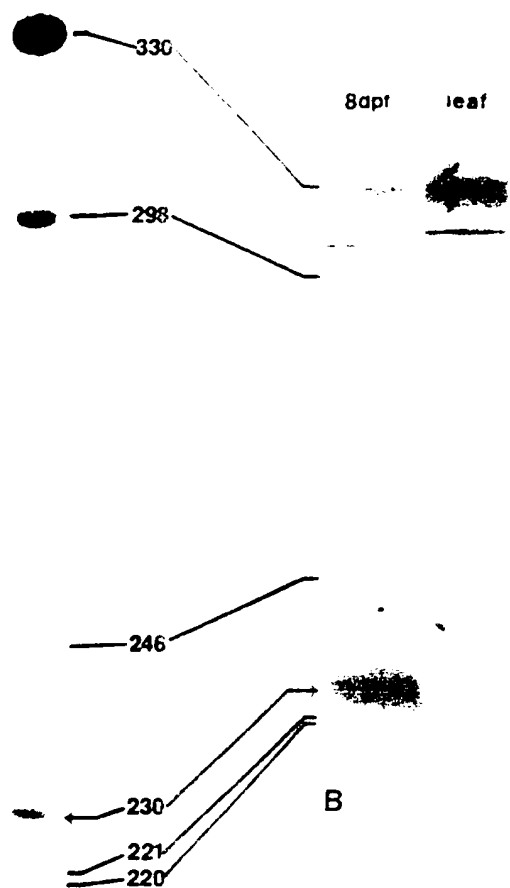


Fig. 3A, B. Nuclease protection of *HaG5* transcription initiation site. **A** Products of mung bean nuclease protection assay separated on 6% sequencing gel. **B** Products of S1 nuclease protection assay separated on 12% sequencing gel. Relative positions of size markers on both gels are indicated by numbers (nucleotides) and lines. The 230 nucleotide fragment protected by embryo RNA is indicated by arrows.

Arginine represents more than 10% of the amino acid residues and along with glutamate (8.2%) accounts for the majority of the charged residues of the mature gene product of *HaG5*. The calculated pI of the predicted *HaG5* gene product is 11.5; therefore, the protein should have a net positive charge at neutral pH. The predicted molecular weight of the mature protein is 17.7 kDa and is in excellent agreement with our estimates from SDS-PAGE (Cohen 1986).

Estimation of sunflower albumin family divergence

HaG5 was isolated by hybridizing a sunflower genomic DNA library with an albumin cDNA probe, *Ha5* (Allen et al. 1987; Cohen 1986). Although *Ha5* does not represent a complete albumin mRNA, it does share sequence homology with *HaG5* over the majority of the transcription unit. Comparison of restriction maps of *HaG5* and *Ha5* suggested these sequences were somewhat divergent (data not shown). The sequence divergence between *HaG5* and *Ha5* is more precisely illustrated in Fig. 4 which shows a compar-

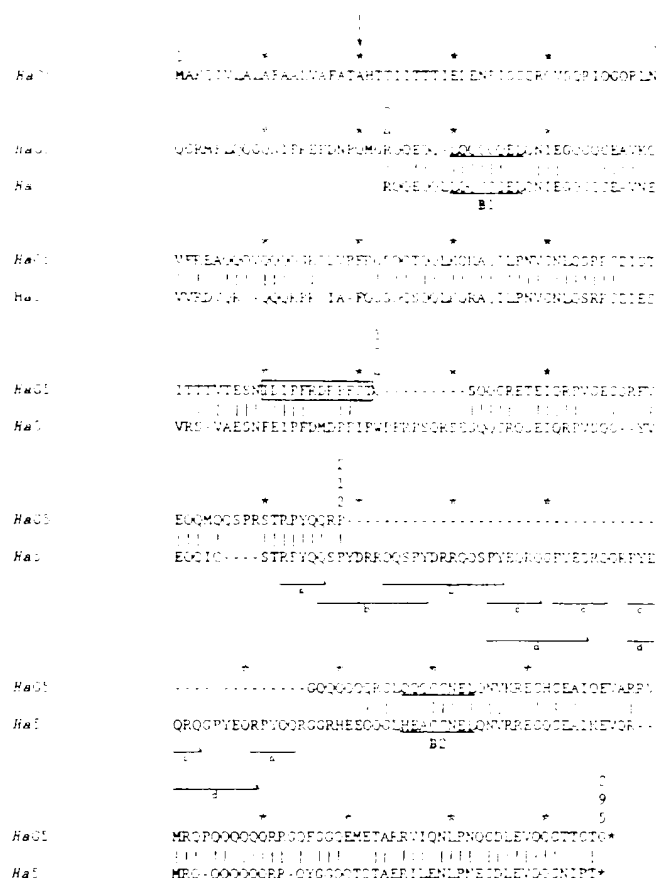


Fig. 4. Comparison of *HaG5* and *Ha5* predicted amino acid sequences. *HaG5* amino acid sequence from Fig. 2 is compared with a predicted amino acid sequence for *Ha5* (Cohen 1986). Gaps were inserted to maximize the homology between the two sequences. Symbols: * homology between indicated residue; \downarrow conservative amino acid change; \downarrow putative cleavage site in hydrophobic leader. *Vertical arrow* indicates putative cleavage site in hydrophobic leader. *Arrows* labeled a, b, and c indicate internal direct repeats in the *Ha5* sequence. *Underlined* regions B1 and B2 indicate homologies with *Bruchid napin* (Gough et al. 1983). *Boxed* sequence is identical to the amino-terminal sequence of the mature sunflower albumin (see Results).

Table 1. Predicted amino acid composition of *HaG5* precursor and mature proteins

Amino acid	Precursor Number (%)	Mature Number (%)	Amino acid	Precursor Number (%)	Mature Number (%)
Ala	14 (4.75)	3 (2.24)	Met	6 (2.03)	2 (2.24)
Asp	16 (5.42)	9 (6.72)	Val	12 (4.17)	4 (2.99)
Asp	5 (1.69)	3 (2.24)	Pro	13 (4.41)	5 (3.97)
Glu	21 (7.12)	11 (8.21)	Gln	71 (24.1)	34 (25.4)
Phe	10 (3.39)	4 (2.99)	Arg	28 (9.49)	17 (12.7)
Gly	17 (5.76)	9 (6.72)	Ser	8 (2.71)	3 (2.24)
His	2 (0.68)	1 (0.75)	Thr	19 (6.44)	7 (5.23)
Ile	16 (5.42)	5 (3.73)	Asn	16 (5.42)	7 (5.23)
Lys	5 (1.69)	1 (0.75)	Tyr	1 (0.34)	1 (0.75)
Leu	15 (5.08)	4 (2.99)	Trp	0 (0.00)	0 (0.00)

ison of the overlapping amino acid sequences predicted from the nucleotide sequences of *HaG5* and *Ha5*. Gaps have been inserted in both sequences to maximize the homology between the two. The most striking feature of Fig. 4

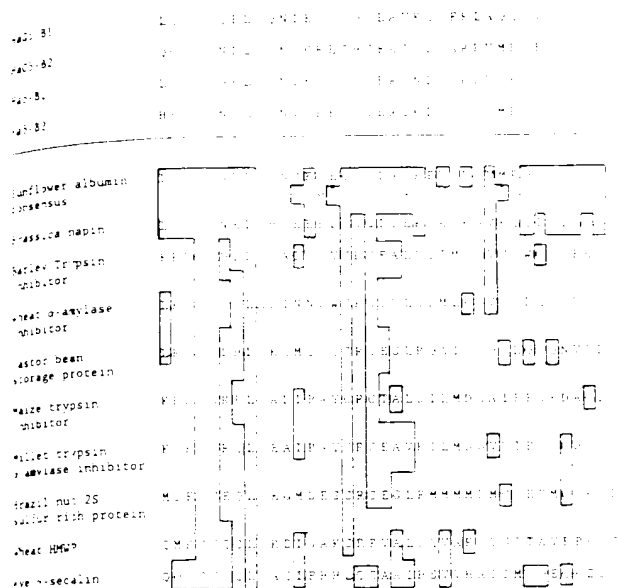


Fig. 5. Phylogenetic conservation of *HaG5* sequences. Sequences sharing homology with *HaG5* "B" regions (Fig. 4) were identified by a computer search of protein sequence data bases or were identified by inspection of sequences compiled by Kreis et al. (1985a, b). Above the line are predicted amino acid (aa) sequences including the B1 and B2 regions of *HaG5* and *Ha5* (see Fig. 4); immediately below the line is a consensus sequence for this region. Sequences for *Brassica napin* 2 (aa 102–136, Crouch et al. 1983), barley trypsin inhibitor (aa 40–75, Odani et al. 1983), wheat α -amylase inhibitor (aa 39–75, Maeda et al. 1983), castor bean storage protein (aa 5–40, Sharief and Li 1982), maize trypsin inhibitor (aa 43–80, Mahoney et al. 1984), millet trypsin α -amylase inhibitor (aa 41–77, Campos and Richardson 1983), Brazil nut 2S sulfur-rich protein, large subunit (aa 9–43, Ampe et al. 1986), wheat high molecular weight prolamins (aa 40–77, Forde et al. 1983) and eye γ -secalin (aa 36–121, Kreis et al. 1985b) are shown below the sunflower albumin consensus sequence and are aligned to maximize homology between the various sequences. Boxes indicate homology with the sunflower albumin consensus sequence.

gins et al. 1987). Processing of the sunflower albumin appears to be most like that observed for castor bean in that it is processed from a rather large precursor polypeptide, but the resulting mature protein is larger and is composed of a single polypeptide containing one or more intra-molecular disulfide linkages (Allen et al. 1987).

A computer search of protein sequence data bases identified significant homologies between the predicted amino acid sequences of *HaG5*, *Ha5* and napin (Crouch et al. 1983). The sequence motif, LQQCCNEI, is represented only once at position 101–109 in the napin precursor, but is found twice in both *HaG5* and *Ha5*. These sequence elements are designated B1 and B2 in Fig. 4. Kreis et al. (1985a, b) defined a storage protein superfamily that included napin as well as other heterogeneous seed proteins. The most significant homologies between the predicted *HaG5* protein and these proteins occur in the peptide domain "B" as defined by Kreis et al. (1985a, b) and include the LQQCCNEI sequence motif. Sequences including the B1 and B2 regions of *HaG5* and *Ha5* were compared with characteristic sequences of this superfamily; the results of these comparisons are summarized in Fig. 5. The most striking observation is the conservation of the LQQCCNEI motif in most sequences and in particular the invariance

of the cysteine residues at the aligned positions 4 and 5 and leucine at position 8. In addition, the cysteine residues at positions 10 and 11 are nearly invariant. The functional significance of the albumin-prolamin superfamily, defined by Kreis et al. (1985a) and further illustrated in Fig. 5, is not clear, however, the striking phylogenetic conservation of these and other sequence motifs (reviewed in Kreis et al. 1985a) suggest a common progenitor for the 2S albumins of dicot and heterogeneous monocot seed proteins including prolamins and various enzyme inhibitors. Particularly relevant to this point are the recent observations of Templeman et al. (1986) that show ostrich fern albumin storage proteins share antigenic determinants and nucleotide sequence homology with *Brassica napin*. Since ferns diverged from the evolutionary line giving rise to angiosperms prior to the divergence of monocots and dicots (Cronequist 1968), these results provide further evidence of the evolutionary relationship between dicot albumins and various monocot seed proteins.

Acknowledgments. This research was supported by grants from the Texas Advanced Technology Research Program and Rhone-Poulenc Associates. RFE was a recipient of a W. R. Grace Foundation Fellowship. We thank Dr. Tom McIntight for his critical review of this manuscript and his help in protein purification and amino acid sequencing.

References

- Aetersfeldt RH, Teplov DB, Hood LB, Kent SBH (1986) High efficiency preparation of proteins from analytical sodium dodecyl sulfate-polyacrylamide gels for direct sequence analysis. *J Biol Chem* 261:4229–4238.
- Allen RD, Nessler CL, Thomas TL (1985) Developmental expression of sunflower 11S storage protein genes. *Plant Mol Biol* 11:167–173.
- Allen RD, Cohen EA, Vonder Haar RA, Orth KA, Ma DP, Nessler CL, Thomas TL (1987) Expression of embryo specific genes in sunflower. In: Davidson EH, Firtel R (eds) *UCLA Symposium on molecular approaches to developmental biology*. Alan R. Liss, New York, p 415.
- Ampe C, Van Damme J, deCastro LAB, Samraio MUAM, Van Montagu M, Vandekerckhove J (1986) The amino acid sequence of the 2S sulfur rich protein from seeds of Brazil nut (*Bertholletia excelsa* H.B.K.). *Eur J Biochem* 159:597–604.
- Benton WD, Davis RW (1977) Screening recombinant clones by hybridization to single plaques in situ. *Science* 196:180–182.
- Borgerwald AG, Lord JM (1983) Egg and *Ruminantia* oviductin subunits are all derived from a single precursor protein. *Eur J Biochem* 127:577–580.
- Campos FAP, Richardson M (1983) The complete amino acid sequence of the bifunctional α -amylase-trypsin inhibitor from seeds of rice (*Oryza sativa* L.). *Plant Physiol* 73:305–308.
- Casey R, Donohue C, Elias S (1986) Legume storage proteins and their genes. Oxford Surv Plant Mol Cell Biol 1:1–95.
- Cohen EA (1986) Analysis of sunflower 2S seed storage protein gene. MS Thesis, Texas A&M University.
- Crone M, Lenhane KM, Simon ME, Ercle R (1983) DNA probe for *Brassica napin* seed storage proteins: evidence from nucleotide sequence analysis that both subunits of napin are provided from a single precursor polypeptide. *J Mol Appl Genet* 2:27–34.
- Cronequist A (1968) The evolution and classification of flowering plants. Houghton Mifflin, Boston, pp 127–131.
- Dair RMK, McGuire BA, Houchens JP (1985) A rapid single-stranded cloning strategy for producing a sequential series of overlapping clones for use in DNA sequencing. Application

- to sequencing the corn mitochondria 1.8S rDNA. *Plasmid* 13:31-40.
- Dalgatarrondo M, Raymond J, Azanza JE (1984) Sunflower seed protein: characterization and storage composition of the globulin fraction. *J Exp Bot* 35:161-168.
- Eriksen ME, Rodin J, Lemman M, Glimelius K, Jørgensen LG, Rask L (1986) Structure of the rapeseed 1.7S storage protein, napin, and its precursor. *J Biol Chem* 261:14376-14381.
- Favard J, Treisman R, Kamen R (1980) Trancription maps of polyoma virus specific RNA. Analysis by two-dimensional nuclease S1 mapping. *Methods Enzymol* 65:718-749.
- Frishaut AM, Lebrach H, Poustka A, Murray N (1983) Lambda replacement vectors carrying polynucleotide sequences. *J Mol Biol* 170:827-842.
- Forde J, Forde BG, Fry RP, Kreis M, Shewry PR, Milfin BJ (1983) Identification of barley and wheat cDNA clones related to the high Mr polypeptides of wheat gluten. *FEBS Lett* 162:360-366.
- Heidecker G, Messing J (1986) Structural analysis of plant genes. *Annu Rev Plant Physiol* 37:439-466.
- Higgins TJV (1984) Synthesis and regulation of major proteins in seeds. *Annu Rev Plant Physiol* 35:191-221.
- Higgins TJV, Chandler PM, Spencer D, Beach LR, Blagrove RJ, Kort AA, Inglis AS (1986) Gene structure, protein structure and regulation of the synthesis of a sulfur-rich protein in pea seeds. *J Biol Chem* 261:11124-11130.
- Higgins TJV, Beach LR, Spencer D, Chandler PM, Randall PJ, Blagrove RJ, Kort AA, Guthrie RF (1987) cDNA and protein sequence of a major pea seed albumin (PA2, Mr ~ 26000). *Plant Mol Biol* 8:37-45.
- Kreis M, Shewry PR, Forde BG, Forde J, Milfin BJ (1985a) Structure and evolution of seed storage proteins and their genes with particular reference to those of wheat, barley and rye. *Oxford Surv Plant Mol Cell Biol* 2:253-317.
- Kreis M, Forde BG, Rahman S, Milfin BJ, Shewry PR (1985b) Molecular evolution of seed storage proteins of barley, rye and wheat. *J Mol Biol* 183:499-502.
- Leach DRF, Stahl FW (1983) Viability of λ phage carrying a perfect palindrome in the absence of recombination nucleases. *Nature* 305:448-451.
- Lord JM (1985) Synthesis and intracellular transport of lectin and storage protein precursors in endosperm of castor bean. *Eur J Biochem* 146:403-409.
- Maeda K, Hase T, Matsubara H (1983) Complete amino acid sequence of an α -amylase inhibitor in wheat kernel. *Biochim Biophys Acta* 743:52-57.
- Mahoney WC, Hermanson MA, Jones B, Powers DD, Cortman RN, Peeck GR (1984) Amino acid sequence and secondary structural analysis of the corn inhibitor of trypsin and activated elastin. *Protein J Biol Chem* 259:8412-8416.
- Mannervik L, Schirmer R, Schlegel J (1975) Nucleotide sequence of the corn α -amylase inhibitor. *Proc Natl Acad Sci USA* 72:1114-1119.
- Maniatis T, Fritsch EF, Sambrook J (1982) *Molecular cloning*. A laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor.
- Messing J (1987) New λ 11 vectors for cloning. *Methods Enzymol* 101:26-78.
- Mount S (1982) A catalogue of splice junction sequences. *Nucleic Acids Res* 10:489-472.
- Odani Y, Linder L, Mott T (1983) The complete amino acid sequence of barley trypsin inhibitor. *J Biol Chem* 258:7998-8003.
- Laroché-Payal M, Dassen J (1986) Identification and characterization of the mPNA for major storage proteins from radish. *Per J Biochem* 157:321-327.
- Sanger F, Coulson AR, Barrell BG, Smith AH, Roe BA (1980) Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *J Mol Biol* 143:161-178.
- Schwenke KD, Pante W, Limow KJ, Schultz M (1979) On oil seed proteins, part II. Identification, chemical composition and some physicochemical properties of the 1.7S globulin (Helianthinin) in sunflower seed. *Die Nahrung* 23:241-254.
- Serling G, Jasin S, Schattner W (1983) Prokaryotes and eukaryotic transcription. *Trends Genet* 1:224-230.
- Shaner ES, Li SL (1982) Amino acid sequence of small and large subunits of seed storage proteins from *Rhynchospora communis*. *J Biol Chem* 257:14753-14759.
- Sun SM, Altenbach SB, Leung FW (1987) Properties, biosynthesis and processing of a sulfur-rich protein in Brazil nut (*Bertholletia excelsa* H.B.K.). *Eur J Biochem* 162:477-483.
- Templeman TS, Demaggio AE, Stein DB (1986) An ostrich fern storage protein shares genetic homology with the 1.7S rapeseed storage protein. *Am J Bot* 73:688.
- von Heintz G (1986) A new method for predicting signal sequence cleavage. *Nucleic Acids Res* 14:4683-4690.
- Youle RJ, Huang AHG (1981) Occurrence of low molecular weight and high cysteine containing albumin storage proteins in oil seeds of diverse species. *Am J Bot* 68:44-48.

Communicated by R.B. Goldberg

Received June 22, 1987

Org
the

Thoma
Steven
Plant
Grad

Summ:
key re-
oid bi-
oid pig-
garris I
syntha-
oid-de-
family
of bear-
some-
with re-
to the
as min-
isopoly-
the rat-
encoded
to both
or infec-
genes
tion of
wound-
eating
activat-
set of C
demon-
nization
in bear-
enzyme
environ-
Key wor-
sequen-

Introduc

A strike
plastic
protect
signifi-
ated fr

* Proc
Diego
** Proc
Environ
USA

Optima



The b-32 protein from maize endosperm, an albumin regulated by the *O2* locus: Nucleic acid (cDNA) and amino acid sequences

N. Di Fonzo¹, H. Hartings², M. Brembilla³, M. Motto¹, C. Soave¹, E. Navarro⁴, J. Palau⁵, W. Rhoads⁶, F. Salamini^{1*}

¹Istituto Sperimentale per la Cerealicoltura, Sezione di Bergamo, Via Stezzano 24, I-24100 Bergamo, Italy

²Università della Basilicata, Potenza, Italy

³Centre d'Investigació i Desenvolupament, CSIC, carrer Girona Salgado, 18, E-08034 Barcelona, Spain

⁴Max-Planck Institut für Züchtungsforschung, D-5000 Köln, Federal Republic of Germany

Summary. The cDNA coding for the b-32 protein, an albumin expressed in maize endosperm cells under the control of the *O2* and *O6* loci, has been cloned and the complete amino acid sequence of the protein derived. A lambda gt11 cDNA library from mRNA of immature maize endosperm was screened for the expression of the b-32 protein using antibodies against the purified protein. One of the positive clones obtained was used to isolate a full-length cDNA clone. By Northern analysis, the size of the b-32 mRNA was estimated to be 1.2 kb. Hybrid-selected translation assays show that the message codes for a protein with an apparent molecular weight of 30–35 kDa. The nucleotide sequence shows that several internal repeats are present. The protein has a length of 303 amino acid residues (mol.wt. 32430 dalton) and its sequence shows the following features: no signal peptide is observable; it contains seven tryptophan residues, an amino acid absent in maize storage proteins; polar and hydrophobic residues are spread along the sequence; several pairs of basic residues are present in the N-terminal region; the secondary structure allows the prediction of two structural domains for the b-32 protein that would fold up giving rise to a globular shape. The cloning of this gene may help in understanding the role of the *O2* and *O6* loci in regulating the deposition of zein, the major storage protein of maize endosperm.

Key words: Zein regulation · *O2* · *O6* · b-32 protein · cDNA cloning

Introduction

The protein b-32 of maize endosperm is a mon-meric albumin with an apparent molecular weight of about 32 kDa, existing in different genotypes in two isoelectric forms: one with a pI of 5.8 and the second with a pI of 6.0. The two variants show similar amino acid composition but minor differences are shown by their tryptic peptide maps. The protein is localized in the soluble part of the cytoplasm and does not bind to any particulate structure (Di Fonzo et al. 1986). Its expression during development is temporary and quantitatively coordinated with the deposition of storage proteins in endosperm tissue (Soave et al. 1981).

In all maize inbreds so far studied the b-32 protein is found, either in the acidic or in the basic form, as a gene product of two codominant alleles; it has also been shown

that the *o2* and *o6* mutants lack this protein (Soave et al. 1981). As both mutants induce a concomitant decrease in the production of zein polypeptides and of protein b-32 it is possible that this protein can act as a *trans*-acting factor regulating storage protein deposition. However, a parallel unrelated control of both zein and b-32 proteins by another gene product(s) cannot be excluded. Whatever the different control mechanisms might be, information on the molecular structure of the b-32 protein may shed some light on its biological role within the endosperm cells.

In this paper we report the isolation and analysis of cDNA clones prepared from mRNA of maize endosperm cells and coding for a product corresponding to the b-32 protein. This has been possible because of the availability of purified anti-b-32 sera (Di Fonzo et al. 1986) for the screening of a lambda gt11 expression library. The complete nucleotide sequence of the b-32 message, as well as the amino acid sequence of the b-32 protein is described.

Materials and methods

Plant material. The wild-type version of the inbred W64A (*Zea mays* L.) was used for large-scale preparation and purification of the basic form of the b-32 protein, as well as for preparing total and poly(A)⁺ RNA. The *o2* and *o6* mutants, in the background of the line W64A, were when needed used to prepare RNA for Northern analysis. In some experiments, wild-type and mutant versions of the maize lines B37 and A66Y were also utilized, as specified in the text. Ears were collected at 25–30 days after pollination, frozen in liquid nitrogen and stored at –80°C until use.

Enzymes and chemicals. DNA restriction endonucleases, DNA polymerase I Klenow fragment, reverse transcriptase and RNase A were purchased from Bethesda Research Laboratories; α -[³²P]dCTP, α -[³⁵S]dATP, L-[³⁵S]methionine and [¹⁴C]-methylated protein mixture were purchased from Amersham International.

Poly(A)⁺ RNA. Total RNA was extracted from dissected endosperms and purified as described by Dean et al. (1985). Poly(A)⁺ RNA was prepared by two cycles of oligod(T) cellulose chromatography (Aviv and Leder 1972).

Expression library in lambda gt11. An expression library was prepared from endosperm poly(A)⁺ RNA, using the cDNA synthesis system from Amersham International. The

synthesized cDNA was size selected (> 300 bp) by agarose gel electrophoresis, and remaining *Eco*RI linkers removed by adsorption on DEAE filters (Whatman DE 81) as described by Dretzen et al. (1981). The *Eco*RI-linked cDNA was ligated to dephosphorylated *Eco*RI-digested lambda gt11 arms (Promega Biotech), and packaged in vitro. Approximately 2×10^7 plaque forming units were obtained, from which 80% were recombinants. The library was amplified on *Escherichia coli* Y1090 (Promega Biotech).

Antibody screening of the λ gt11 library. Serum for screening was raised in rabbits and purified as described by Soave et al. (1981). The library was plated and after incubation at 42°C for 4 h the plates were overlaid with dry nitrocellulose filters saturated with 10 mM isopropyl β -D-thiogalactopyranoside, and further incubated at 37°C for 3 h (Young and Davis 1983). After this second incubation, filters were washed with saturation buffer (PBS: 2% bovine serum albumin; 0.05% Nonidet NP40). PBS was 10 mM phosphate buffer, pH 7.5; 150 mM NaCl. The serum was diluted with the saturation buffer (1:100) and used for incubating the filters at 37°C overnight. After recovering the serum, filters were washed with a solution of 10 mM phosphate buffer, pH 7.5; 1 M NaCl; 0.05% Nonidet NP40 for 1 h at room temperature and incubated for 2 h at room temperature with 125 I-protein A (> 30 mCi/mg, Amersham International) in saturation buffer (at 5×10^5 cpm/ml). Positive clones were purified by successive cycles of antibody screening, until all phages in a plate showed a positive signal.

In vitro translation and immunoprecipitation. Immunoprecipitation of in vitro translation products was performed as described by Davis et al. (1986). Proteins were analyzed by SDS-12% polyacrylamide gel electrophoresis.

Northern blot analysis. One microgram of poly(A)⁺ RNA was resolved by electrophoresis on a formaldehyde-agarose gel (1.3% agarose; 2.2 M formaldehyde; 1% 3-[N-morpholino]propanesulfonic acid). The gel was soaked in $20 \times$ SSC for 30 min and the RNA transferred to nitrocellulose filters. $1 \times$ SSC = 15 mM sodium citrate, pH 7.0; 150 mM NaCl. The filter was hybridized according to Maniatis et al. (1982).

Hybrid-selected translation. Denatured DNA (1 μ g) was spotted on nitrocellulose filters with the aid of a Minifold (Schleicher and Schüll). Filters were washed with $4 \times$ SSC and baked at 80°C under vacuum. The filters were prehybridized in 68% formamide; 10 mM piperazine-N,N'-bis(2-ethanesulfonic acid); 0.4 M NaCl, pH 6.4; 700 μ g/ml poly(A)⁺ RNA for 1 h at 52°C. Poly(A)⁺ RNAs (30 μ g) were hybridized for 3 h in 120 μ l of the above buffer except for poly(A)⁺ RNA at 52°C. Filters were then washed five times with 10 mM Tris-HCl; 2 mM EDTA; 0.5% SDS, pH 8.0, and five times with 10 mM Tris-HCl; 2 mM EDTA, pH 8.0. Bound RNA was eluted at 55°, 75° and 95°C in 200 μ l of H₂O in 2 mM EDTA and quenched on ice. Carrier tRNA from calf liver (10 μ g/ml) and 3 M sodium acetate (pH 5.6; 20 μ l) were added. The samples were both precipitated and washed with 70% ethanol and the pellets used to direct protein synthesis in the rabbit reticulocyte lysate.

Restriction endonuclease mapping. Restriction endonuclease cleavage sites were determined by single or double digests

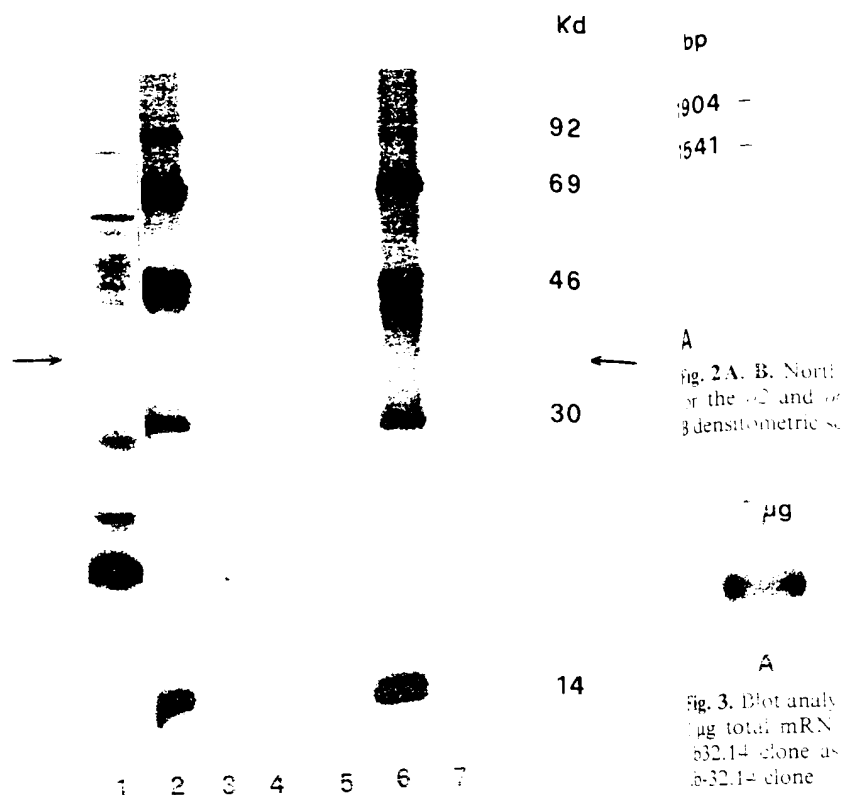


Fig. 1. The electrophoretic pattern of in vitro synthesis directed by poly(A)⁺ RNA (1 μ g) extracted from the inbred line A69Y wild-type is shown in lane 1. The position of migration of b-32 polypeptide (arrow) corresponds to that of purified b-32 (lane 7). Lanes 3, 4 and 5 correspond to immunoprecipitated products from in vitro translated poly(A)⁺ RNA from wild-type, o2 and ob, respectively. Lanes 2 and 6 were loaded with a standard set of [14 C]-labelled proteins.

with various restriction endonucleases. Digestion products were resolved in conventional horizontal agarose gels.

DNA sequencing. The dideoxynucleotide chain termination method of Sanger et al. (1977) was followed using the bacteriophage vectors M13mp18 and M13mp19.

Computer analysis. The hydrophobicity plot of the deduced amino acid sequence was obtained according to Kyte and Doolittle (1982). The prediction of the b-32 secondary structure was made according to the procedures of Garnier et al. (1978) and Chou and Fasman (1974) in the computer version of Parrilla et al. (1986).

Results

Control of b-32 messages by the o2 and ob loci

Previous results (Soave et al. 1981; Di Fonzo et al. 1986) have shown the absence of the b-32 polypeptide in protein extracts of the maize endosperm mutants o2 and ob. Here we have studied to what extent b-32 mRNA is present in the two mutants (Fig. 1). Lane 1 displays the patterns of the in vitro protein synthesis primed by total poly(A)⁺ RNA extracted from the wild-type endosperms in the background of the inbred line A69Y. Lanes 2 and 6 were loaded with molecular weight markers, while lane 7 shows the posi-



Fig. 3. Northern blot analysis of total mRNA from a b-32.14 clone and a b-32.14- clone. The position of the band is indicated by an arrow.

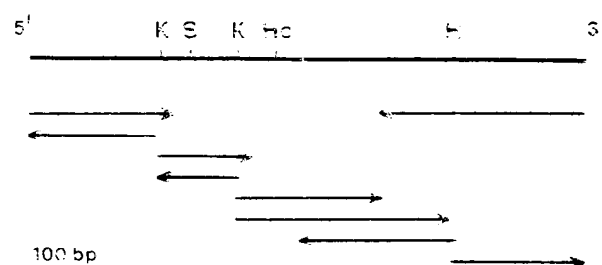
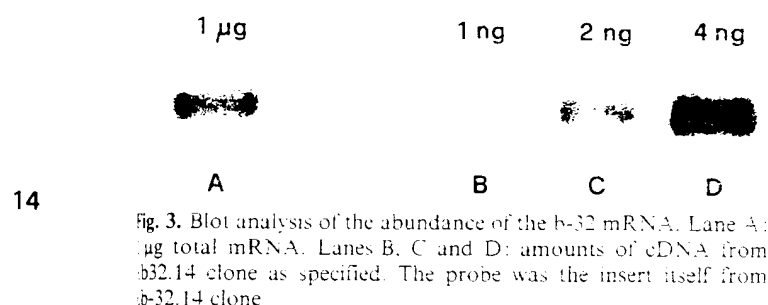
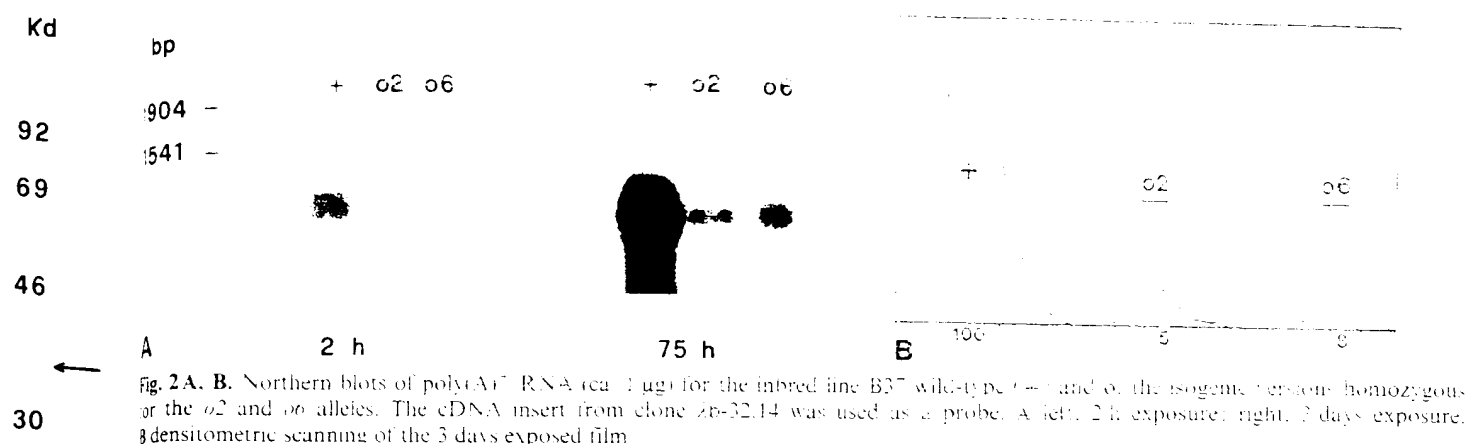


Fig. 5. Restriction endonuclease map of the cDNA insert of zb-32.66 and sequencing strategy. The directions of sequencing of each restriction site are indicated by arrows. The endonucleases are shown as: K, *Kpn*I; S, *Sma*II; Hc, *Hinc*II; H, *Hind*III.

tion of dansylated purified b-32 (arrows). A major in vitro product corresponds to this position in lane 1. Lanes 3, 4 and 5 show for the wild-type *o2* and *o6* mRNA extracts the in vitro translation products precipitated by an anti-b-32 antiserum. The in vitro synthesis of the b-32 product is detectable only for the wild-type, confirming previous conclusions on the role of *O2* and *O6* loci in the control of protein b-32 in the cells of the endosperm. It can also be observed in lanes 3 and 5, corresponding to the wild-type and *o6* extracts, a precipitation of relatively small quantities of zein-type proteins. This finding is probably due to the exceptionally large amount of this zein message in maize endosperms.

cDNA cloning and immunodetection of b-32 clones

A lambda gt11 expression library was prepared from endosperm mRNA isolated 20 days after pollination. An anti-b-32 antiserum was used to isolate cDNA inserts expressing the b-32 polypeptide. Approximately 1.5×10^5 recombinant phages were analyzed by filter hybridization and various clones showing positive signals were isolated. Six of these clones were further analyzed in detail and purified by replating and screening with the b-32 antiserum. Only clones designated zb-32.14 and zb-32.19 were confirmed positive. Their cDNA inserts have a size of 0.7 and 0.5 kb, respectively, as shown by *Eco*RI digestion and subsequent gel electrophoresis. The cDNA insert from clone zb-32.14 was amplified and used as a probe for Northern blot and hybrid-selected translation experiments.

synthesis directed
nbred line A69Y
migration of b-32
ated purified b-32
precipitated prod-
om wild-type, *o2*
with a standard

estion products
arose gels.

uin termination
using the bacte-

of the deduced
ig to Kyte and
b-32 secondary
ures of Garnier
n the computer

oci

zo et al. 1986)
ptide in protein
2 and *o6*. Here
A is present in
he patterns of
otal poly(A)⁺
ns in the back-
6 were loaded
shows the posi-

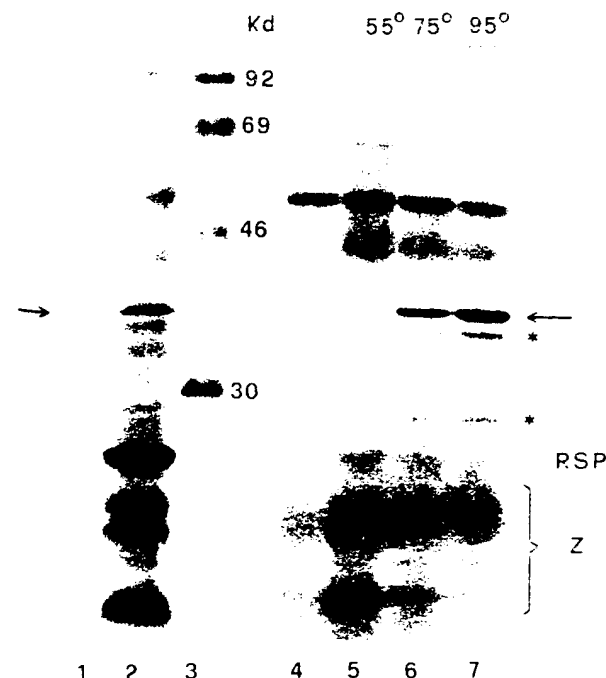


Fig. 4. Hybrid selected translation experiment. The cDNA from zb-32.14 clone was used as a probe for hybridization. Lane 1, immunoprecipitate of an in vitro translation of 1 µg poly(A)⁺ RNA extracted from B37 wild-type. Arrows indicate position of polypeptide b-32. Lane 2, the same but not immunoprecipitated. Lane 3, standard set of molecular weight markers. Lane 4, endogenous translation products of the rabbit reticulocyte lysate. Lanes 5-7, hybrid selected mRNAs translated after post-hybridization washes at increasing temperatures (55°, 75° and 95° C, respectively). Position occupied by zeins (Z), glutelin-2(RSP) and minor components (*) are indicated.

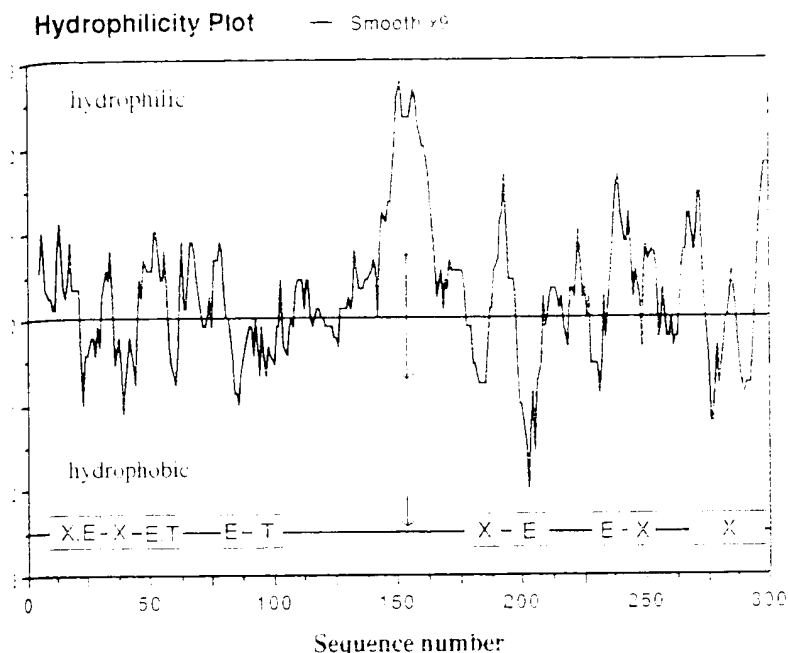


Fig. 7. Top: hydrophilicity plot of the deduced amino acid sequence of protein b-32. The programme of Kyte and Doolittle (1982) was used and the values of each position were plotted against the residue number of protein b-32. Bottom: prediction of α -helix (H), β -extended structures (E) and reverse turn (T) for the b-32 protein, following and combining the procedures of Chou and Fasman (1974) and Garner et al. (1978).

This value corresponds to a protein consisting of about 300 amino acid residues. The abundance of the b-32 message in poly(A)⁺ RNA extracts of wild-type endosperms was determined by blot analysis, using for comparisons increasing amounts of cDNA insert from λ b-32.14 (Fig. 3). The results show that there are 2–3 copies of b-32 mRNA per every 10³ copies of total mRNA.

Hybrid selected translation experiments using the DNA insert from λ b-32.14 as a probe are presented in Fig. 4. A number of appropriate controls were performed: lane 1 indicates the position in a gel of the protein precipitated with the b-32 antibody from in vitro translation of poly(A)⁺ RNA from wild-type B37 endosperm; in the in vitro pattern of total poly(A)⁺ RNA (lane 2) the major band of 32 kDa is present; lane 4 shows the polypeptide bands corresponding to endogenous translation products of the rabbit reticulocyte lysate. The hybrid-selected translation samples occupy lanes 5–7. Post-hybridization washings of filters with bound mRNAs were carried out at 55°, 75° and 95° C. The hybrid-selected products were then electrophoretically run as shown in the figure. At the lowest temperature, unspecifically hybridized mRNA was eluted and the bound messages mainly gave rise on translation to zeins, glutelin-2 (RSP protein), a protein diffusing in the gel around a position corresponding to a molecular weight of 35 kDa and the endogenous polypeptides of the lysate. On raising the washing temperature, the pattern of translated proteins gradually changes. The polypeptide band of 32 kDa, immunoprecipitable with anti-b-32 antibody, strongly increases in intensity, whereas the other translation products observed with washing at 55° C gradually disappear. In addition to the b-32 mRNA, two minor components of lower molecular weight than b-32 polypeptide (as marked in the figure) seem to be still bound at 95° C. Their level, however, is by far lower than that of b-32 mRNA.

Sequence of a full length b-32 mRNA cDNA clone

The cDNA insert from the phage λ b-32.14 was used as a probe to rescreen the library in order to identify a full-

length cDNA clone. The rescreening yielded 40 positive clones of different insert lengths. The largest clone (λ b-32.66) showed a cDNA insert of about 1 kb which was considered to correspond to a full-length b-32 cDNA clone.

The restriction map of the cDNA insert from λ b-32.66 is shown in Fig. 5. Its sequence was determined by the strategy also depicted in Fig. 5. Figure 6 shows the nucleotide sequence as well as the amino acid sequence of the protein encoded by the largest open reading frame. The insert of clone λ b-32.66 corresponds to the expected full-length b-32 cDNA clone. There is an open reading frame of 909 nucleotides, corresponding to a protein of 303 amino acid residues. The translational start codon is preceded by a TGA stop codon that would invalidate the translation of any larger polypeptide. The 3' flanking region contains a typical polyadenylation signal located 47 nucleotides downstream from the stop codon. The sequence of the cDNA clone λ b-32.14 was also obtained. The length of λ b-32.14 was 662 bp and, within the coding region, its sequence was different from the full-length cDNA at position 305 (substitution of A by G).

Structural analysis of the b-32 polypeptide

The molecular weight of the 303 residues polypeptide as deduced from the sequence of the λ b-32.66 clone is 32430 dalton, which is in good agreement with values determined by SDS-gel electrophoresis for the b-32 protein. In addition, no sequence with the characteristics of a signal peptide is observable after the start codon.

In the second half of the translatable region a number of nucleotide duplications are present, including from: 469 to 480 a GAA GAA GAA GAA sequence (i.e. Glu-Glu-Glu-Glu); from 493 to 507 a GGA GGA GGA GGA GGT (i.e. Gly-Gly-Gly-Gly-Gly); from 508 to 525 a GCA GAT GCA GAT GCA GAT (i.e. Ala-Asp-Ala-Asp-Ala-Asp); from 544 to 564 a GCG GCG GCG GCG GCG GCG GCT (i.e. Ala-Ala-Ala-Ala-Ala-Ala-Ala); from 589 to 606 a AAG CTG GTG AAG CTG GTG (i.e. Lys-Leu-Val-Lys-Leu-Val); from 856 to 873 a GCC GTC GCC GCT GCC

derlined; the
species under
loci, as is
circumstantial
responds to

ern blotting
cb (Fig. 2).

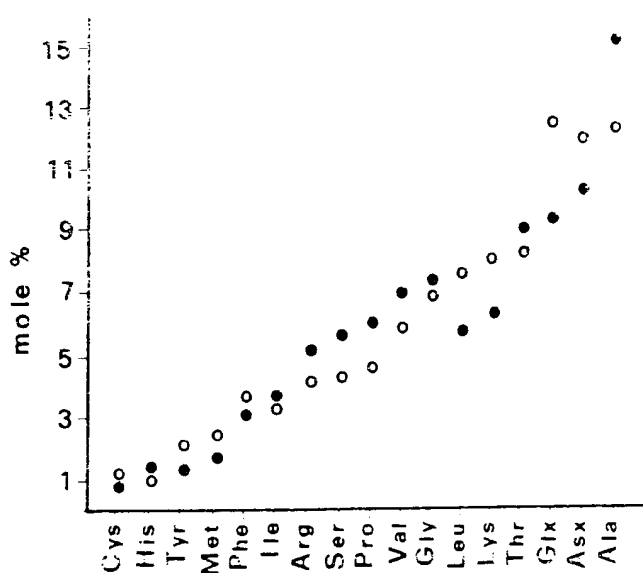


Fig. 8. cDNA-based amino acid composition (closed circles) compared to the one chemically determined (open circles) as reported in the paper by Di Fonzo et al. (1986).

GCT (i.e. Ala-Ala-Ala-Ala-Ala-Ala); and from 891 to 903 a GAC AAC GAC GAT GAC (i.e. Asp-Asn-Asp-Asp-Asp). An inverted repeat of 9 nucleotides (from 732 to 751) is also observed in the same region.

With respect to the amino acid sequence of this protein, there are different features that deserve attention. Polar and hydrophobic residues are spread along the whole chain. The molecule can be divided approximately in two. The extreme N-terminal region (residues 1-70) shows an enrichment in proportion of pairs of basic residues. The C-terminal domain is rich in repeats, either of the same residue or of groups of two or three residues. To obtain more information concerning the two postulated domains of the molecule, some predictions were made of its secondary structure (Fig. 7). The upper part of the figure shows the hydrophilicity plot of the polypeptide chain. It can be observed that, within the N- and C-terminal domains of the b-32 protein, hydrophobic and hydrophilic segments alternate. A small zone divides the two regions around residue 160: the zone corresponds to a highly hydrophilic sequence very rich in acidic residues (6 out of 7 are Glu or Asp) that should be flexible and located at the surface of the b-32 protein molecule. To make predictions of the b-32 secondary structure two procedures were followed. The structure obtained with both procedures coincide for most of the segments with compact secondary structures. The lower part of Fig. 7 shows the predicted alpha, beta and turn structures along the b-32 polypeptide chain. One can also observe the existence of a central region, probably poorly structured, separating the N- and C-terminal domains that are rich in secondary structure motifs. These two regions have all the requirements to fold up, giving rise to two well defined structural domains of the molecule.

Discussion

The results presented in this paper strongly indicate that the cloning strategy adopted was successful in isolating

cDNA sequences containing an open reading frame coding for protein b-32. In particular it has been shown that: (1) the clones isolated select a mRNA coding for a protein of the expected size; (2) this protein is correctly recognized by an anti-b-32 antiserum; and (3) the b-32-specific mRNA level observed in a Northern blot experiment was very low in the *o2* and *o6* mutants as expected based on the absence of b-32 protein in these genotypes.

The amino acid composition derived from the sequence shows a good similarity, although not a perfect coincidence, with that determined for the purified b-32 protein (Di Fonzo et al. 1986; Fig. 8). The differences noted for few amino acids are easily explained by those artifacts inherent in the chemical determination of amino acid content, such as level of purity of the protein and differential losses of amino acids during acid hydrolysis. In the protein b-32, 2.0% tryptophan was found, a value which is in contrast to the lack of this amino acid in zein storage proteins (Mossé et al. 1966).

Following the folding pattern revealed by the structural analysis of the b-32 deduced sequence, the b-32 protein appears to be a typical globular protein. Its level in developing maize endosperm (Soave et al. 1981) is in the range of an average value for messages coding for endosperm albumins and globulins. Despite the relative abundance of the protein, we believe it may play a direct regulatory role on zein synthesis. Based on genetic evidence, the b-32 protein was credited with such a positive regulatory role in zein deposition (Soave et al. 1981; Di Fonzo et al. 1986). Further studies may substantiate this assumption and, particularly, could reveal if, as postulated, the b-32 polypeptide is actually the gene product of the *O6* locus.

Acknowledgements: This work was supported by EEC contract number BAP-0214-1(A) in the framework of the Biotechnology Action programme, by Ministero dell'Agricoltura e delle Foreste, Italy, special grant "Tecnologie Avanzate in Agricoltura", and by Fundación Ramón Arece. The authors thank Prof. R. Farias for the help in preparing the expression library, and Drs. E. Querol and A. Parrilla for doing a great part of the computer analysis of this work. J.P. acknowledges to NATO a fellowship for a sabbatical stay at the Istituto Sperimentale per la Cerealicoltura, Sezione di Bergamo.

References

- Aviv H, Leder P (1972) Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid-cellulose. *Proc Natl Acad Sci USA* 69:1408-1412
- Chou PY, Fasman GD (1974) Prediction of protein conformation. *Biochemistry* 13:211-245
- Davis LG, Dibner MD, Battey JF (1986) *Basic methods in Molecular Biology*. Elsevier, New York
- Dean C, Van Den Elzen P, Tamaki S, Dunsmuir P, Bedbrook J (1985) Differential expression of the eight genes of the petunia ribulose biphosphate carboxylase small subunit multi-gene family. *EMBO J* 4:3055-3061
- Di Fonzo N, Manzocchi L, Salamini F, Soave G (1986) Purification and properties of an endospermic protein of maize associated with the Opaque-2 and Opaque-6 genes. *Planta* 167:587-594
- Dretzen G, Bellard M, Sassone-Corsi P, Chambon P (1981) A reliable method for the recovery of DNA fragments from agarose and acrylamide gels. *Anal Biochem* 112:295-298
- Garnier J, Dsguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* 120:97-120

ste J. Doolittle, hydrophobic amino acids. T. F. a laboratory, New York. Mossé J. Baud, protéines de acides aminés, de 8:331-344. Parrilla A. Don, program for ropathic seg

- time coding
in that: (1)
a protein
recognized
specific mRNA
is very low
in the absence
- the sequence
coincidence,
protein (Di
ed for few
its inherent
content, such
the losses of
protein b-32,
in contrast
ge proteins
- the structural
b-32 protein
el in devel-
the range
endosperm
undance of
ulatory role
b-32 pro-
ory role in
t al. 1986).
n and, par-
olypeptide
- EC contract
otechnology
elle Foreste.
tura", and
f. R. Farias
s. E. Querol
iter analysis
or a sabbat-
ura. Sezione
- ective globin
ndylic acid-
nformation
s in Molecu-
2. Bedbrook
f the petunia
multi-gene
Purification
e associated
7:587-594
p (1981) A
s from agar-
28
of the accu-
ting the sec-
20:97-120
- Wate J. Doolittle RF (1982) A simple method for displaying the hydrophobic character of a protein. *J Mol Biol* 157:105-132
- Maniatis T, Fritsch EF, Sambrook J (1982) *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory Press, New York
- Mossé J, Baudet J, Landry J, Moureaux T (1966) Etude sur les protéines du maïs. II. Comparaison entre les compositions en acides aminés et les proportions mutuelles des fractions protéiques de grains normaux et mutants. *Ann Physiol Veg* 8:331-344
- Arrilla A, Domenech A, Querol E (1986) A Pascal microcomputer program for prediction of protein secondary structure and hydrophobic segments. *Computer Appl Biosciences* 2:221-215
- Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74:5463-5467
- Soave C, Tardani L, Di Fonzo N, Salamini F (1981) Zein level in maize endosperm depends on a protein under the control of the opaque-2 and opaque-6 loci. *Cell* 27:405-410
- Young RA, Davis RW (1983) Yeast RNA polymerase II genes: isolation with antibody probes. *Science* 222:778-782

Communicated by H. Saedler

Received January 18, 1988

Structure of a Gene Encoding the 1.7 S Storage Protein, Napin, from *Brassica napus**

(Received for publication, December 22, 1986)

Lars-Göran Josefsson, Marit Lenman, Mats L. Ericson, and Lars Rask

From the Department of Cell Research, The Swedish University of Agricultural Sciences, Box 596, S-751 24 Uppsala, Sweden

A rapeseed chromosomal region containing a gene (*napA*), which encodes the 1.7 S seed storage protein (napin), was isolated in several overlapping recombinant clones from a phage λ genomic library. Following restriction enzyme mapping of the genomic region, a subclone containing the *napA* coding region as well as some 1.1 and 1.4 kilobases of DNA from the 5' and 3' regions, respectively, was mapped and sequenced. The gene turned out to lack introns. Southern blotting analyses utilizing a napin cDNA clone as a probe revealed the presence of on the order of 10 napin genes in the rapeseed genome. The major polyadenylated transcript encoded by these genes was shown to be an 850-nucleotide species, the initiation site of which was mapped onto the *napA* gene. The major initiation site for transcription is located some 33 nucleotides downstream from a sequence perfectly conforming to the consensus sequence of a TATA box. Further analyses of the sequence revealed several features that may be of relevance for the expression of the napin genes.

Napin, or the 1.7 S protein, is one of the major seed storage proteins in *Brassica napus*. It is expressed in a tissue-specific manner, apparently under the influence of abscisic acid (Crouch and Sussex, 1981; Crouch *et al.*, 1983). The mature protein, which is rather basic, consists of two subunit polypeptides that are linked by disulfide bridges (Ericson *et al.*, 1986; Lönnnerdal and Janson, 1972). Comparison of amino acid sequences of the subunits with the sequence of a cDNA clone has shown that the initial translation product, a 20-kDa precursor, contains both the subunit polypeptides as well as polypeptide stretches that are removed during the maturation of the protein (Ericson *et al.*, 1986). By analogy with other storage proteins, the final product is thought to reside in specialized organelles, protein bodies, within the seed cells (Larkins and Hurkman, 1978). As far as is known, the sole function of napin is to serve as a nutrient source during germination and initial development of the seedling. Confirmatory evidence that napin, like other storage proteins, possesses minor heterogeneities in the amino acid sequence stems from protein separation data (Lönnnerdal and Janson, 1972) as well as protein sequencing (Ericson *et al.*, 1986) and the analysis of cDNA clones (Crouch *et al.*, 1983; Ericson *et al.*,

1986). As an initial step toward an increased understanding of the regulation of napin genes, we have isolated and sequenced a member of what turns out to be a small gene family.

MATERIALS AND METHODS AND RESULTS¹

DISCUSSION

We have isolated and sequenced a gene encoding napin. The gene is a member of a small family with some 10 genes. Transcription of an as yet unknown number of these genes yields an 850-nucleotide-long mRNA, the cap site of which was mapped onto the *napA* sequence. We have compared our sequence with that of another napin gene, pGNA, as well as with previously sequenced cDNA clones (Crouch *et al.*, 1983; Ericson *et al.*, 1986). The *napA* sequence is completely identical to the pNAP1 cDNA clone that we have previously sequenced (Ericson *et al.*, 1986). This makes us rather confident that we have sequenced an expressed copy of the napin gene family, although we have no formal proof that this is the case.

Comparison with the pGNA gene sequence revealed that, apart from single nucleotide changes, a quite frequently occurring divergence in the coding region is insertions of one or two triplets in pGNA relative to *napA*. These occur in four and two instances, respectively (data not shown). Apart from one previously reported triplet deletion in the pN1 cDNA clone (Crouch *et al.*, 1983). These are the first examples of differences that affect the length of the primary sequence of the translated napin product. The number of nucleotide changes in the coding region is also higher when comparing *napA* with pGNA than with any of the previously sequenced cDNA clones (data not shown). It is interesting to speculate whether these observations may be related to the fact that *B. napus* is an amphidiploid of *Brassica campestris* and *Brassica oleracea*. It might be expected that the genes derived from one of the respective parental species would be more homologous to each other than when comparing across the parental border. We are presently attempting to assign parentship of isolated napin genes by comparison with Southern blots of genomic DNA from the three species. Preliminary data² indicate that the *napA* gene most likely is derived from *B. oleracea*.

* This work was supported by The Swedish Research Council for Natural Sciences, The Swedish Research Council for Forestry and Agriculture, and the Stiftelsen Brinkgården. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EMBL Data Bank with accession number(s) J02798.

¹ Portions of this paper (including "Materials and Methods," "Results," and Figs. 3 and 4) are presented in miniprint at the end of this paper. The abbreviations used are: SDS, sodium dodecyl sulfate; kb, kilobase; dNTP, deoxynucleotide triphosphate; AMV, avian myeloblastosis virus; hn, heterogenous nuclear. Miniprint is easily read with the aid of a standard magnifying glass. Full-size photocopies are available from the Journal of Biological Chemistry, 9650 Rockville Pike, Bethesda, MD 20814. Request Document No. 86 M-4366, cite the authors, and include a check or money order for \$3.20 per set of photocopies. Full size photocopies are also included in the microfilm edition of the Journal that is available from Waverly Press.

² M. L. Ericson, unpublished data.

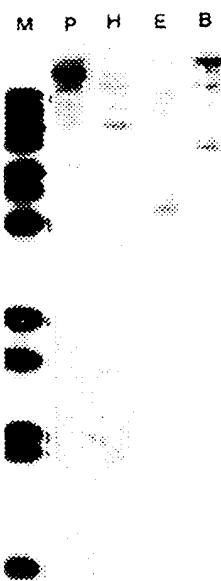


FIG. 1. Genomic restriction fragments hybridizing with napin cDNA sequences. Genomic DNA was cut with restriction enzymes. The generated fragments were separated and blotted onto nitrocellulose filters as described under "Materials and Methods." Nick-translated pNAP1 cDNA was used as a probe in hybridization to these filters. The enzymes used were *B. BamHI*; *E. EcoRI*; *H. HindIII*, and *P. PvuII*. The size marker (*M*) used was an end-labeled *BstEII* digest of phage λ DNA. Sizes of the marker bands were (from top to bottom): 8454, 7242, 6369, 5687, 4822, 4324, 3675, 2323, 1929, 1571, 1264, and 702 base pairs.

M R



FIG. 2. Northern blotting and hybridization of rapeseed mRNA to pNAP1 cDNA. mRNA was purified and separated on denaturing agarose gels as described under "Materials and Methods." After transfer to nitrocellulose filters the immobilized mRNA was hybridized to a nick-translated cDNA probe. *R* denotes the RNA lane; *M*, the marker lane. The marker used was a denatured *HinfI* digest of pBR322. The autoradiogram reveals the marker bands hybridizing to nick-translated pUC19. The sizes of the bands are 1631 and 517/506 nucleotides, respectively.

A C G T R

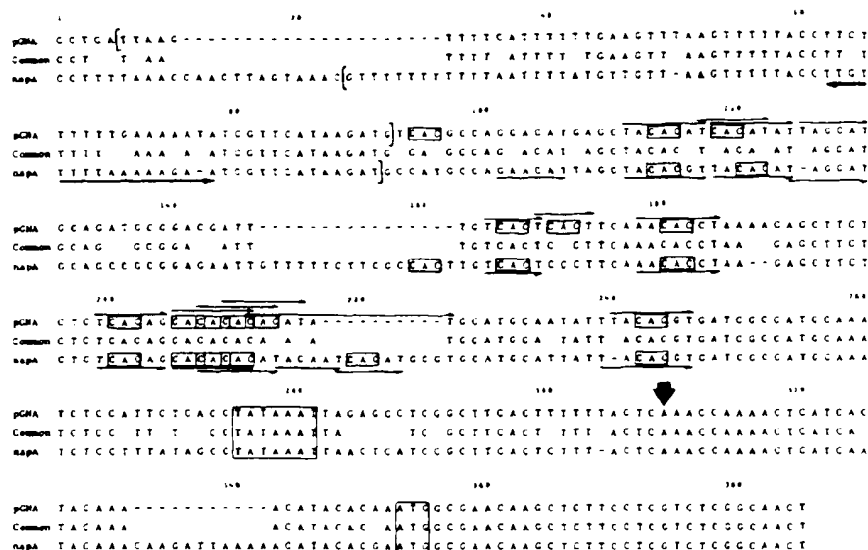


FIG. 5. Transcript cap site mapping of napin mRNA. An 18-mer oligonucleotide, complementary to a napin sequence just downstream from the initiation codon, was synthesized. This synthetic oligonucleotide, ^{32}P end-labeled and unlabeled in the respective cases, was annealed to either mRNA or M13 DNA covering this region on the minus strand. In separate reactions the primer was allowed to be elongated to the 5' end of the napin transcripts or to prime a standard set of sequencing reactions. The products were separated on a gradient sequencing gel. Lane *R* shows the terminated forms that were elongated on the mRNA. Lanes *A*, *C*, *G*, and *T*, the respective sequencing reactions.

With regard to the primary translation product, comparisons of all the known sequences have made us aware of an interesting repeated structure in the removed parts of the napin polypeptide. All of the previously sequenced cDNA clones and the two genomic clones discussed here conform to this structure. It consists of a stretch of 7 or 8 amino acids, $X-X\cdots(-)X$, where X denotes hydrophobic and $-$ negatively charged amino acids, respectively. These sequences in *napA* are shown boxed in Fig. 6. The negatively charged amino acid in brackets is only present in the first copy of the repeat which occurs in the amino-terminal part of the precursor sequence, before the small subunit. The second copy of the repeat occurs within the removed sequence which is present between the small and large subunits. These two repeats in fact carry almost all of the negative charges that are contained in the processed parts of the precursor (Ericson *et al.*, 1986). It is possible that these repeats are involved in processes relevant for the translocation, intracellular transport, and/or deposition of napin into protein bodies. Alternatively, they could serve as signals in the proteolytic processing steps necessary for the generation of mature napin. However, confirmation of a possible role of these repeats in the above processes will have to await experiments directly aimed at these points.

We have noted several interesting features in the sequence of *napA* (and *pGNA*) that may be of relevance to different aspects of gene regulation. It is tempting to speculate that the 5' hairpin region and the TACACAT repeat region may be directly involved in the transcriptional activation of the gene and that the 3' hairpin region may be involved in the termination of transcription. There is ample precedence in the literature for the former point, i.e. degenerate (or non-degenerate) repeats as well as alterations in DNA topology (possibly manifesting itself in cruciform structures) have been implied in gene regulation in several systems (Gidoni *et al.*, 1985; Hall *et al.*, 1982; Harland *et al.*, 1983; Serfling *et al.*, 1985). It appears more doubtful what role hairpin loops may play in

FIG. 7. Alignment of the *napA* promoter region and the promoter region of the pGNA napin gene. The nucleotide sequences of the promoter regions of *napA* and the pGNA napin gene were aligned by use of the ALIGN program (Dayhoff *et al.*, 1979) run with the UN matrix, a break penalty of 2 and 100 random runs. CAC trinucleotides are boxed and perfect or degenerate versions of the TACACAT repeats are indicated by arrows. The TATA box and initiation ATG are boxed for reference. The major transcription cap site is indicated by an arrow. Brackets at the 5' end encompass sequences with a tendency to form hair-pin loops.



Acknowledgment—Dr. Steve R. Scofield is gratefully acknowledged for making his sequence of the pGNA napin gene available to us prior to publication.

REFERENCES

- Bankier, A. T., and Barrell, B. G. (1983) in *Techniques in Nucleic Acid Biochemistry* (Flavell, R. A., ed) Vol. 85, pp. 1-73, Elsevier Scientific Publishers Ltd., Limerick, Ireland.
- Birnstiel, M., Busslinger, M., and Strub, K. (1985) *Cell* **41**, 349-359.
- Breathnach, R., and Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349-383.
- Courey, A. J., and Wang, J. C. (1983) *Cell* **33**, 817-829.
- Crouch, M. L., and Sussex, I. M. (1981) *Planta (Berl.)* **153**, 64-74.
- Crouch, M. L., Tenbarger, K. M., Simon, A. E., and Ferl, R. (1983) *J. Mol. Appl. Genet.* **2**, 273-283.
- Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1979) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. D., ed) Vol. 5, Suppl. 3, pp. 345-362, National Biomedical Research Foundation, Washington, D. C.
- Ericson, M. L., Rödin, J., Lenman, M., Glimelius, K., Josefsson, L.-G., and Rask, L. (1986) *J. Biol. Chem.* **261**, 14576-14581.
- Frischauf, A.-M., Lehrach, H., Poustka, A., and Murray, N. (1983) *J. Mol. Biol.* **170**, 827-842.
- Gidoni, D., Kadonaga, J. T., Barrera-Saldana, H., Takahashi, K., Chambon, P., and Tjian, R. (1985) *Science* **230**, 511-517.
- Hall, B. D., Clarkson, S. G., and Tocchini-Valentini, G. (1982) *Cell* **29**, 3-5.
- Harland, R. M., Weintraub, H., and McKnight, S. L. (1983) *Nature* **302**, 38-43.
- Hentschel, C. C., and Birnstiel, M. L. (1981) *Cell* **25**, 301-313.
- Hohn, B. (1979) *Methods Enzymol.* **68**, 299-309.
- Kamlen, H., Scheil, J., and Kreuzaler, F. (1986) *EMBO J.* **5**, 1-8.
- Larkins, B. A., and Hurkman, W. J. (1978) *Plant Physiol.* **62**, 256-263.
- Lönnnerdal, B., and Janson, J.-C. (1972) *Biochim. Biophys. Acta* **278**, 175-183.
- Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
- Morelli, G., Nagy, F., Fraley, R. T., Rogers, S. G., and Chua, N.-H. (1985) *Nature* **315**, 200-204.
- Mizuuchi, K., Mizuuchi, M., and Gellert, M. (1982) *J. Mol. Biol.* **156**, 229-243.
- Proudfoot, N. J., and Brownlee, G. G. (1976) *Nature* **263**, 211-214.
- Rackwitz, H.-R., Zehetner, G., Frischauf, A.-M., and Lehrach, H. (1984) *Gene (Amst.)* **30**, 195-200.
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463-5467.
- Serfling, E., Jasin, M., and Schaffner, W. (1985) *Trends Genet.* **1**, 224-230.
- Simon, A. E., Tenbarger, K. M., Scofield, S. R., Finkelstein, R. R., and Crouch, M. L. (1985) *Plant Mol. Biol.* **5**, 191-201.
- Sinden, R. R., Carlson, J. O., and Pettijohn, D. E. (1980) *Cell* **21**, 773-783.
- Staden, R. (1980) *Nucleic Acids Res.* **8**, 3673-3694.
- Staden, R. (1982) *Nucleic Acids Res.* **10**, 4731-4751.
- Weier, H., König, M., and Gruss, P. (1983) *Science* **219**, 626-631.
- Yanisch-Perron, C., Vieira, J., and Messing, J. (1985) *Gene (Amst.)* **33**, 103-119.

Continued on next page.

Supplementary material to
Structure of a Gene Encoding the 1.7S Storage Protein, Napin,
from *Brassica napus*

by

Josefsson, L.-G., Lennan, M., Ericson, M. L., and Rasm, L.

MATERIALS AND METHODS

Plants

B. napus seeds of a dihaploid variety of "Svenska Karol" were generously provided by Dr. Lena Bengtsson, Svalöv AB, Sweden. This repeated line was used throughout these studies.

Isolation of DNA

100 g quantities of etiolated, frozen leaf tissue were homogenized along with solid CO_2 in a Waring blender. When the powder was starting to thaw, 100 ml of 40 mM Tris-HCl, pH 8.0/10 mM EDTA/1% SDS (sodium dodecyl sulphate) were added and the suspension incubated at 60°C for 20 min with gentle agitation. This was followed by two gentle extractions with a 2:1 mixture of chloroform:isoamylalcohol. The aqueous phase was retained and the DNA precipitated. The precipitate was collected by centrifugation, rinsed with 70% ethanol, dried lightly and resuspended in 8 ml of TE (10 mM Tris-HCl, pH 8.0/1 mM EDTA). RNA was degraded by an incubation for 30 minutes at 67°C after the addition of RNase A to a final concentration of 100 $\mu\text{g}/\text{ml}$. 0.1 volumes of a solution containing 50 mM Tris-HCl, pH 7.5/4 M EDTA/0.5% SDS/1 mg/ml proteinase K were then added and the mixture incubated at 50°C for 30 minutes. The mixture was then extensively dialysed against TE. After dialysis the DNA was extracted twice with chloroform:isoamylalcohol and then precipitated. The DNA was rinsed with 70% ethanol, lightly dried and resuspended in TE. This procedure yielded easily restrictable DNA with a mean size of some 10 kb as determined by low voltage electrophoresis in a 2% agarose gel.

Southern blotting

10 μg portions of rapeseed DNA were digested to completion with different restriction enzymes and a vector control was run with the TBE (Tris/Borate/EDTA) buffer system (Maniatis et al. 1982). After light staining with ethidium bromide the gel was immersed in 0.5 M NaCl for 5 minutes. After the denaturation the DNA in the gel was transferred and transferred to nitrocellulose filters as described (Maniatis et al. 1982). The subsequent treatment of the filters was also according to Maniatis et al. (1982).

Isolation of mRNA and Northern blotting

mRNA was isolated as described by Eilstrup et al. (1984). Denaturing agarose gels were prepared and run according to Maniatis et al. (1982). 2 μg of denatured mRNA were loaded on a 1% agarose formaldehyde gel and subjected to electrophoresis. Transfer of the mRNA to nitrocellulose filters and the subsequent treatment of the filters was according to standard procedures (Maniatis et al. 1982).

Nick-translation and hybridization to Southern blots

0.1-0.2 μg portions of pNAP1 cDNA were nick-translated to obtain radiolabelled probe. Prehybridizations and hybridizations were done with formamide-containing solutions according to standard protocols (Maniatis et al. 1982). Washing of filters was done at high stringency, i.e. 3 mM sodium citrate/0.1 M NaCl/0.5% SDS at 65°C, 5 times 1 h. Filters were exposed on X-ray film with intensifying screens at -70°C.

Construction of genomic library and screening for napin clones

Rapeseed DNA (120 μg) was partially degraded with MboI under conditions that predominantly yielded fragments in the size range of 2-5 kb. DNA molecules of this size class were further purified by fractionation on 5-40% sucrose gradients in 1 M NaCl that were run for 8 h at 19,000 rpm in a Beckman SW40 rotor. The fractions containing 15-25 kb DNA were pooled. The DNA was precipitated, resuspended and phosphatase treated to further reduce the risk of insert concatamerization during ligation. After removal of the phosphatase by extraction the DNA was precipitated, pelleted, rinsed and resuspended in TE to a concentration of 0.5 $\mu\text{g}/\mu\text{l}$. The pNAP1 vector DNA was double cleaved with BamHI and EcoRI and the small linker-sequences removed by isopropanol precipitation of the DNA (Prachauf et al. 1983). The BamHI and EcoRI fragments were ligated and packaged in λ extracts for packaging of phage lambda in vitro were prepared according to Honn (1979).

Conditions of ligation and packaging of lambda particles in vitro were investigated on a small scale prior to preparation of the library. The conditions finally chosen for the library construction were the following: 4 μg of vector and 2.5 μg of 0.1-2 kb insert DNA were ligated and packaged in vitro after dividing into 10 mixes containing each: 25 μl buffer A (20 mM Tris-HCl, pH 8.0/5 mM MgCl₂/0.5% 8-mercaptoethanol/1 mM EDTA), 1.65 μg DNA (in 10 μl), 5 μl of mix (6 mM Tris-HCl, pH 8.0/5 mM spermidine/18 mM MgCl₂/5 mM ATP/0.1% 8-mercaptoethanol), 17.5 μl sonication extract and 25 μl freeze/thaw lysate. This yielded a total of 2.2×10^8 plaque forming units (pfu). The library was subsequently applied on 10 large screening plates of 2x12 cm. Some 2×10^5 clones of the amplified library were screened by spreading 1×10^4 pfu on each screening plate. Replica nitrocellulose filters were prepared and pretreated as described (Maniatis et al. 1982). Conditions for hybridization are given above. Recombinant phages were purified by two consecutive screenings on regular 10 cm LB/agar plates. Growth of recombinant phages and purification of phage DNA were according to established protocols (Maniatis et al. 1982).

Mapping of genomic clones and subcloning

Lambda recombinant clones were mapped by combining the procedure of Bacterius et al. (1984) with a set of complete digestions with either SalI alone or with SalI together with either of six other restriction enzymes. Southern blots from gels on which the latter digestions were analysed were prepared and hybridized with labelled pNAP1 cDNA. Subcloning of a fragment containing the napin gene was done by purification of the fragment on an agarose gel cast with low melting temperature agarose. The purified fragment was subcloned into pUC19 (Vanish-Peterson et al. 1985). The subclone was mapped by conventional digestion/double digestion techniques (Maniatis et al. 1982).

Nucleotide sequencing

Nucleotide sequencing was performed according to Sanger et al. (1977) with [³²S] dGTP as the labelled nucleotide. The M13 vectors used were mp18 and mp19 (Vanish-Peterson et al. 1985). Both the shotgun procedure (Barker and Barker, 1983) and directed subcloning into M13 of fragments derived by restriction enzyme digestion were used. The sequence data was read by use of the DB system (Staden, 1982, Staden, 1982).

Mapping of the napin transcription start site

An 18-mer oligonucleotide, 5'-AGCAAGAGCTTCTCCG-3' was [³²P] end-labelled with polynucleotide kinase (Maniatis et al. 1982). Approximately 0.1 μg of rapeseed poly(A)⁺ RNA of the labelled oligonucleotide were added to 1 μg of RNA in a 10 μl mix

that in addition contained: 35 u Human placental RNase inhibitor, 18 mM Tris-HCl, pH 8.3 measured at 42°C, 25 mM NaCl and 8 mM MgCl₂. After annealing for 1 h at room temperature, unlabelled mRNAs to a final concentration of 200 μM each and 1.5 units of AMV reverse transcriptase were added. The sample was then incubated at 42°C for 20 min and subsequently treated as a regular sequencing gel sample. Approximately 1 μl of the mixture (5000 cpm) was loaded onto the gel and run alongside a reference set of sequencing reactions.

Databases

The three major data bases: EMBL, GDB and GENBANK were used in the sequence comparisons.

RESULTS

Southern and Northern blotting analyses

As an initial step towards defining the complexity of the rapeseed genome with regard to napin genes we decided to use pNAP1, a cDNA clone which encodes napin (Ericson et al. 1985), as a radioactive probe in Southern blotting analyses. 10 μg portions of total rapeseed DNA were in separate reactions digested to completion with four different restriction enzymes. Following separation of the generated DNA fragments on agarose gels the fragments were denatured and transferred to nitrocellulose filters. Hybridization to the filters of nick-translated pNAP1 cDNA yielded the pattern shown in Figure 1. The different enzymes yielded between 8 and 13 hybridizing bands. Since it is not known to what extent the enzymes may cut within individual napin genes, there is no way of deducing an exact gene number. Nevertheless, considering the data as a whole it appears reasonable to assume that there are in the order of 10 genes for napin. Now many of these hybridizing bands that represent expressed napin genes is at present not clear.

Irrespective of the fact that several genes may be expressing napin, one well defined, major napin mRNA species was evident when rapeseed embryonal mRNA was subjected to Northern blotting with the cDNA probe (Figure 2). In addition to the major 850 nucleotide transcript, a diffuse population of RNA species is also evident. This range in size from approximately 900 to 1500 nucleotides, and as a whole constitutes quite a significant fraction of the total hybridizing material. We cannot at present determine whether these larger mRNAs represent a differentially expressed or differently polyadenylated species of napin transcripts or simply are contaminating mRNA which has not yet been polyadenylated. In light of the fact that polyadenylation appears to be a much more site specific process compared to that of transcriptional termination (Bastin et al. 1985) we favour the latter explanation.

Isolation and restriction mapping of napin genomic clones

A genomic phage library was constructed with DNA from a dihaploid line of *B. napus*. Screening of 1.1×10^5 recombinants with the pNAP1 cDNA clone as the probe yielded eight positive clones. DNA was prepared from these clones after they had been purified by two consecutive screenings. Mapping of the genomic clones showed that four of the positive recombinants were overlapping clones containing the same gene, which we have designated napin. Figure 3 displays the restriction map of this region, as well as the individual clones that cover the region. A 3 kb HindIII-BglII fragment hybridized to the pNAP1 cDNA probe was subcloned into plasmid pUC19 (Vanish-Peterson et al. 1985) and further mapped by conventional techniques (Maniatis et al. 1982). Figure 4 shows the map that was obtained and a comparison with the pNAP1 cDNA restriction map.

It has been shown in other plant gene systems that the cis signals involved in regulating transcriptional initiation usually are contained within sequences that are located reasonably close to the transcribed part of the gene (Rasmussen et al. 1986; Morelli et al. 1985). Thus, we considered it likely that all the linked sequences involved in transcriptional regulation were contained in this subclone and consequently decided to sequence the whole insert of the subclone.

Sequencing of the napin gene

The entire sequence of the 3.3 kb fragment was determined in overlapping sequence reactions on both strands by a combination of "shotgun" sequencing and sequencing of individual restriction endonuclease (R) subclones. Both the universal 11-mer sequencing primer and synthetic oligonucleotides (8-mers) complementary to sequences within the subclones were used to obtain the complete sequence. The sequencing strategy is represented in a schematic fashion below the restriction map in Figure 4. This represents a minimal estimate of sequence data that were collected. Sequences that were well represented in the "shotgun" clones, the transcribed region in particular, were determined with higher frequency than is apparent from the figure. In addition, many individual reactions were performed more than once.

Mapping of the initiation site for transcription

The transcription cap-site of napin mRNA was determined by mRNA directed primer extension. A synthetic 33-nucleotide complementary to mRNA sequences close to the initiation ATG was [³²P] end-labelled, annealed to mRNA and subsequently elongated to the 5' end of napin mRNA by the incorporation of unlabelled nucleotides mediated by AMV reverse transcriptase. Figure 5 shows the elongated and terminated primer alongside the sequence reactions obtained by letting the same oligonucleotide, unlabelled in this case, prime sequencing reactions on an M13 shotgun clone that covered this region on the minus strand. When mapped onto the sequence of the napin gene the major initiation site is at the A in position 1102. The minor bands correspond to positions 1098, 1112 and 1113. Thus, the major site of transcriptional initiation appears to be located 33 nucleotides downstream from a sequence which conforms to the consensus of a TATA box see below.

General features of the sequence

Figure 6 shows the sequence of the 1295 nucleotides of the HindIII-BglII subclone insert. The related sequence of the coding region is also shown above the nucleotide sequence in one letter code. The sequence that is contained in the pNAP1 cDNA clone (Ericson et al. 1984) is shown within brackets. It is absolutely identical to that of napin. In this context it is worth noting that the pNAP1 cDNA clone was isolated from a cDNA library constructed from the same dihaploid rapeseed line that was used for these studies. A thick arrow indicates the major transcription start site. This is preceded by an encoded TATA containing sequence (Brenthach and Chaboud, 1981). A dotted line shows an imperfect CAT box (Brenthach and Chaboud, 1981) which is 11 bp at nt 1114. The sequence of the poly(A) tail is shown in the context of the coding region one poly(A) addition signal (Proudfoot and Brownlee, 1976) is found (underlined) at actual site where the poly(A) tail is added, as deduced from a comparison with the pN1 and pN2 cDNA clones (Frouch et al. 1983). Figure 6 also shows a second set of TATA/poly A

addition signals (boxed/underlined) at nucleotides 2453 and 2931, respectively. We presently do not know whether this part of the sequence represents an expressed portion of the genome. Considering the size of a hypothetical transcript and the relative positions of ATG's and termination codons within it, we think it is less likely that this sequence is expressed (at least at the protein level). However, we are presently testing this point by a direct examination. On the opposite strand there are several TATA boxes and polyA addition signals. An A/T-rich sequence between positions 1078-1117 would on the opposite strand correspond to a closely spaced polyA addition signals. One additional polyA addition signal occurs at position 1963 (plus strand numbering). Towards the 3' end of the minus strand three TATA boxes occur at 495, 751 and 783 (plus strand numbering), the two first of which are part of a 14 bp direct repeat. Slightly further downstream on the minus strand are two additional polyA addition signals 106 and 188 (plus strand numbering). Although we seriously doubt whether any of the above sequences constitute functional signals, we can at present not strictly rule it out.

Hairpins, repeats and palindromes

The major direct and inverted repeats of the sequence are indicated by arrows, pairwise connected as indicated by the lettering. Solid arrows indicate perfect, dotted arrows imperfect repeats. The E repeat has been observed and discussed previously (Ericson *et al.*, 1986). In addition to the repeats shown in Figure 4, the region 1078-1117 has several overlapping direct repeats. A two-headed arrow indicates a slightly imperfect palindromic sequence present some 210 nucleotides upstream of the TATA box. Two different regions, nucleotides 824-859 and 2139-2208, display features which appear to result in quite a strong tendency to form hairpins (and possibly cruciform structures). Several other regions in the sequence exhibit some tendency to form hairpins. However, the unique feature of the regions that are discussed here is that within a rather short stretch of nucleotides (60-70) several different hairpin structures can be generated simply by sliding the hairpin arms relative to each other. The above sequences are able to form 6 and 7 different hairpin structures, respectively, with between 9 and 15 base pairs in the stem (data not shown). With regard to the former of these sequences a corresponding region in the pDNA napin gene (S. Scofield, personal communication) is, although the sequence differs somewhat from napA, able to form 5 different alternative structures with between 9 and 13 base pairs in the stem. In addition, these regions in napA and in the pDNA gene are both very A/T-rich. This strengthens the suggestion that the sequences may be involved in a local perturbation of the DNA structure. The sequences involved are shown within brackets in Figure 7. Both of the sequences in the napA gene are rather suggestively placed. One (positions 474-485) upstream and the other (positions 2139-2208) downstream from the transcribed part of the gene. No doubt, the possibility to form several alternative hairpins could be of importance in stabilising a non B-DNA structure, particularly if the regions are under negative superhelical stress (Mizuuchi *et al.*, 1982). However, it has been argued that the kinetics of cruciform formation may restrict the importance of such reactions *in vivo* (Cousy and Varg, 1983). The question whether transcriptional activation of eukaryotic chromatin is at all influenced by torsional stress *in vivo* is also controversial. Data favouring both opinions have been put forward (Harland *et al.*, 1983; Sinden *et al.*, 1980). Another interesting feature of the promoter region is that within a sequence 10-140 base pairs upstream of the TATA box the trinucleotide CAC is repeated 11 times. Most copies of the CAC trinucleotide occur as part of a tandemly repeated, degenerate heptamer, which in turn is repeated 12 times. Figure 7 shows the region of interest with the sequences of napA and the pDNA napin gene aligned to maximize the homologies. The CAC trinucleotides and the degenerate repeats are indicated in the figure. The consensus sequence of the repeat considering all the different copies is TACACAT. The TATA box and initiation ATG are boxed for reference. The major cap-site is also indicated by an arrow.

Comparison with other nucleotide sequences

A search with the napA 5' region sequences against the three major data bases as well as against recently published (and not yet entered) sequences of some storage protein genes from other species failed to reveal any features that we could tentatively identify as being related to gene regulation. We were also unable to find sequences in napA related to SV40 enhancer core sequences (Weiner *et al.*, 1983) unless allowing for 1 or more mismatches.

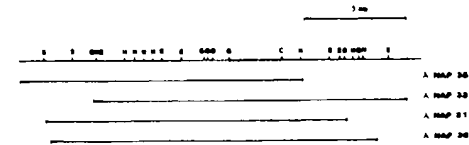


Figure 3: Restriction map of the genomic region containing the napA gene. Individual lambda recombinant clones were mapped as described in Materials and Methods. The figure shows the map of the genomic region and the parts contained in different recombinants. The measuring bar corresponds to 5 kb of DNA. The enzymes used were AclI; ClaI; EcoRI; HindIII; KpnI; MspI; PstI; SalI; XbaI. The hatched area indicates the part that hybridized to pnapA1.

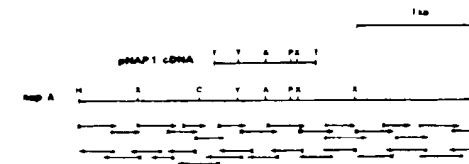


Figure 4: Restriction map of napA and sequencing strategy. The 3.3 kb HindIII - BglII subclone in pUC19 was mapped with conventional techniques. The figure shows the map obtained for the insert and how it compared to the map previously obtained for pnapA1 cDNA. Measuring bar corresponds to 1 kb of DNA. The enzymes used were: AclI; ClaI; EcoRI; HindIII; KpnI; MspI; PstI; SalI; XbaI and YmaI. Below the map is a schematic representation of the sequencing strategy, as discussed in the text. I denotes reactions primed by the universal 17-mer primer on either shotgun clones or restriction enzyme derived (RI) subclones. e denotes reactions primed by synthetic 18-mer primers within different subclones.

Nucleotide Sequence of a Member of the Napin Storage Protein Family from *Brassica napus**

(Received for publication, December 19, 1986)

Steven R. Scofield† and Martha L. Crouch

From the Department of Biology, Indiana University, Bloomington, Indiana 47405

We have begun the molecular characterization of genes encoding napin, the 1.7 S embryo-specific storage protein of *Brassica napus*. Genomic Southern blot analysis indicates that napin is encoded by a multigene family comprised of a minimum of 16 genes. Two DNA fragments containing single napin genes have been recovered from *B. napus* genomic libraries. We have determined the nucleotide sequence of one member of the napin gene family, gNa. The gene has a simple structure lacking introns and containing the canonical features expected for genes transcribed by RNA polymerase II. The site of the initiation of transcription was determined to be 37 base pairs upstream of the initiation codon by S1 and primer extension analyses. A gene-specific hybridization probe from the 3' non-translated portion of gNa was used to demonstrate transcription of gNa.

As the sequences of seed proteins from different plants become known, homologies between proteins with drastically different properties are being detected. For example, several of the diverse 2 S proteins found in seeds have been shown to share sequence homology: the methionine-rich Brazil nut storage protein,¹ the allergenic storage protein in castor bean endosperm (Sharief and Li, 1982), the very basic 1.7 S storage protein in rapeseed embryos (Crouch *et al.*, 1983), and a trypsin inhibitor from barley (Odani *et al.*, 1983). Also, these proteins are related to the prolamin storage proteins such as γ -secalin from rye (Kreis *et al.*, 1985) and α -gliadin from wheat (Kasadara *et al.*, 1984), even though the prolamins are much larger and are hydrophobic rather than hydrophilic. In many cases, the properties of the specific proteins are the result of repeated sequences that differ between them (Higgins, 1984). Despite the different physical properties conferred by these repeats, all of the proteins accumulate to high levels during seed development, are stored during the period of developmental arrest separating embryogeny from germination, and are then degraded during seedling growth. Thus, the basic pattern of temporal expression has been retained. This class of storage proteins is particularly important for animal nutrition, since they usually have higher levels of the sulfur-

containing amino acids than the other abundant seed proteins (Youle and Huang, 1981).

We have been studying the expression of the genes for the 1.7 S storage proteins from *Brassica napus* L. (rapeseed), the napins. Using a cloned cDNA probe from one of the napin family members, transcripts can first be detected early in embryo development, just after the major tissue systems have been delineated (Crouch *et al.*, 1985). Levels of napin mRNA increase until they constitute about 8% of the total mRNA at the end of cell division,² stay high for 15 days, and then decrease to barely detectable levels in dry seeds. Napin transcripts cannot be detected at any other time in development. However, this pattern of expression reflects the average of several napin genes. In order to study regulation of napin gene expression in detail, it is necessary to analyze family members individually.

In this paper, we begin an analysis of the napin gene family by determining the minimum number of napin genes and by cloning and sequencing one member of the family. From S1 protection and primer extension experiments, we have determined where in the sequence transcription begins and that this family member is expressed.

MATERIALS AND METHODS³

RESULTS

Napin Gene Family—It is clear from genomic Southern blots that napin is encoded by a family of genes. At least 14 fragments, ranging from 2 to 23 kb⁴ in size, hybridize with different intensities to a napin cDNA probe pN1 when genomic DNA is restricted with *EcoRI* (Fig. 1A). *EcoRI* does not cleave within any cloned napin sequence. The hybridization pattern observed is the same whether the probed DNA is made from a single plant or from a population, indicating that this pattern is not due to population polymorphism (data not shown). The hybridization pattern is also unchanged when probes representing the 5' and 3' halves of the pN1 coding sequence are tested, indicating that all the bands are due to homology with the napin coding sequence and not a repeated sequence in one portion of the cDNA clone pN1 (data not shown).

Fig. 1B is a genomic reconstruction experiment. The genomic clone λ BnNa, described later, was digested with *EcoRI*,

* This work was supported in part by National Science Foundation Grant PCM-83-16403 (to M. L. C.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EMBL Data Bank with accession number(s) J02782.

† Recipient of a Floyd Memorial Fellowship. Present address: Dept. of Molecular Genetics, Plant Breeding Institute, Cambridge CB2 2LQ, Great Britain.

¹ S. Sun, personal communications.

² A. J. DeLisle and M. L. Crouch, unpublished data.

³ Portions of this paper (including "Materials and Methods" and Figs. 2 and 3) are presented in miniprint at the end of this paper. Miniprint is easily read with the aid of a standard magnifying glass. Full size photocopies are available from the Journal of Biological Chemistry, 9650 Rockville Pike, Bethesda, MD 20814. Request Document No. 86M4334, cite the authors, and include a check or money order for \$3.20 per set of photocopies. Full size photocopies are also included in the microfilm edition of the Journal that is available from Waverly Press.

⁴ The abbreviations used are: kb, kilobase(s); bp, base pair(s).

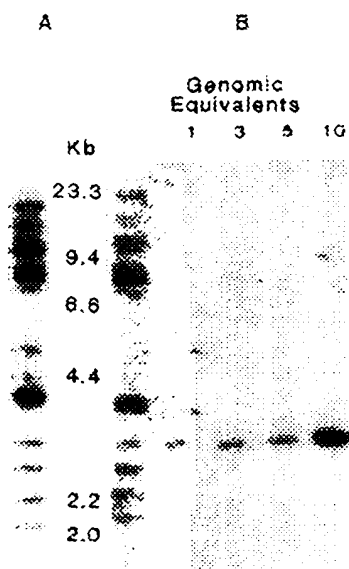


FIG. 1. 4, genomic Southern blot of an *Eco*RI digest of *B. napus* DNA probed with nick-translated pN1 and washed at $T_m - 22^\circ\text{C}$. Data have been placed by single copy signals; 2 designates signals with intensity corresponding to two copies. *B.*, genomic reconstruction: Lane 1, 10 μg of *B. napus* DNA digested with *Eco*RI; Lanes 3–5, *ABnA* DNA digested with *Eco*RI and loaded to simulate 1, 3, 5, and 10 haploid genome equivalents based on 1.6 pg/haploid *B. napus* genome (Verma and Rees, 1974). The filter was probed with nick-translated pN1.

and dilutions representing 1, 3, 5, and 10 copies/haploid genome were electrophoresed beside *Eco*RI-digested genomic DNA. We conclude that the fragments which have the least intense signals contain single napin genes, and the stronger signals represent two or more genes. By this analysis there are at least 16 napin genes/haploid genome. The more intense signals result either from fragments of similar size that contain single genes or linkage of two or more napin genes on an *Eco*RI fragment.

Isolation of Genomic Napin Clones—A genomic library was constructed in the λ vector EMBL4 from *B. napus* DNA digested partially with *Sau*3A. Two unique napin genomic clones, designated λ BnNa and λ BnNb, were isolated when 4×10^6 recombinant phage were screened by plaque hybridization with a nick-translated pN1 napin cDNA probe (Crouch *et al.*, 1983).

The napin genomic clones were analyzed by restriction nuclease mapping and Southern blot hybridizations. Each phage contains just one napin gene, and only the napin gene region hybridizes to cDNA made from embryo RNA, indicating that no other abundant embryo transcripts are encoded by the cloned DNA (data not shown). Comparison of restriction maps derived for genomic napin subclones with those of the cDNA clones pN1 and pN2 shows that these genes do not encode the messages represented by the cDNA clones (Fig. 2). ABnNa was chosen for more thorough examination.

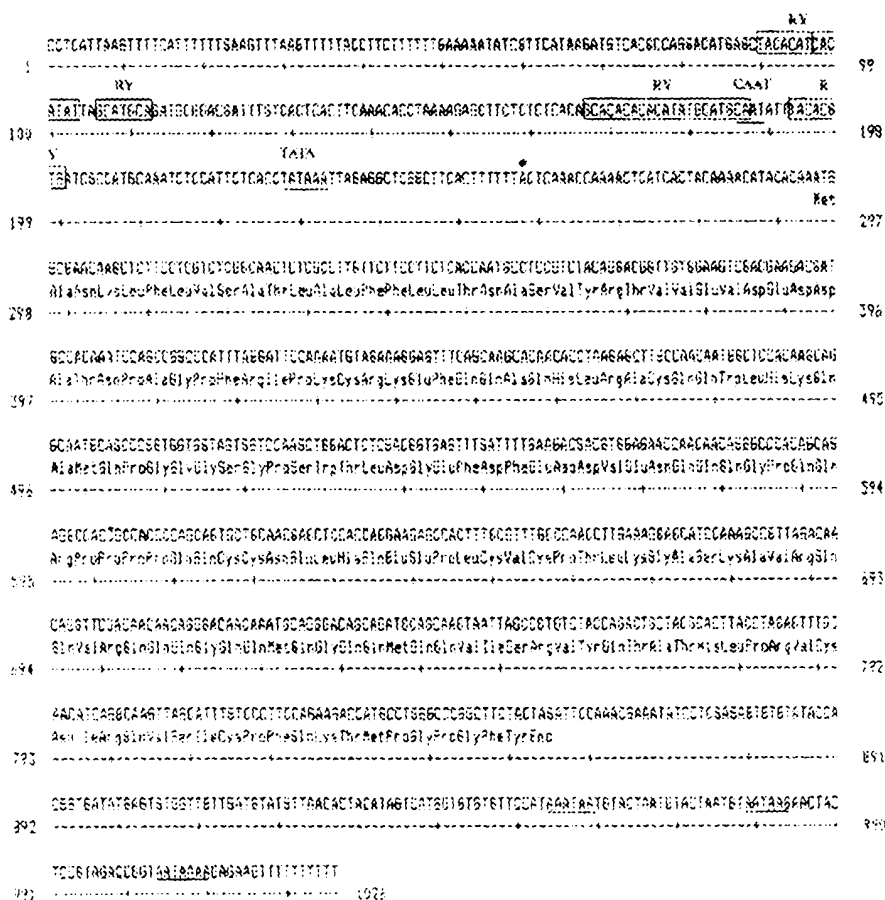


FIG. 4. Nucleotide sequence and deduced amino acid sequence of gNa. The boxed sequences labeled RV are alternating purine-pyrimidine elements 7 bp or longer. Also labeled are: CAAT sequence at 186; the TATA box at 228; the primer extension mapped cap site at 259 (*italic*); the initiating ATG at 295; and three sequences with homology to the consensus polyadenylation processing sequence at 954, 979, and 1004 (*underlined*).

been sequenced by the method of Maxam and Gilbert (Maxam and Gilbert, 1980). This sequence is comprised of 561 coding bp, 252 5' and 172 3' flanking bp (Fig. 5).

The napin reading frame is the only open reading frame of significant length on either strand. The 5' end of this sequence is very AT rich (64%) and is marked by many blocks of 4-6 consecutive A or T residues. A TATA box closely matching the consensus is found 70 bp upstream from the ATG codon initiating the napin precursor. This is the first ATG codon downstream of the TATA sequence. Forty-two bp upstream of the TATA box is the sequence CAAT (position 196, Fig. 4). Though in the expected position, this sequence shows only 4 bp of homology to the 9 bp consensus element shown to be important for efficient promoter recognition (Hancot *et al.*, 1980). Three regions of alternating purine-pyrimidine residues occur upstream of the TATA box: between positions 90 and 110 are three 7 bp alternating purine-pyrimidine units, at position 167 a block of 11 consecutive purine-pyrimidine residues occurs, and at position 192 an 8-bp unit is found.

The 3' untranslated region is high in AT content (67%). Plant genes frequently are found to contain multiple sequences resembling the consensus element associated with polyadenylation of mRNA (Fitzgerald and Sherk, 1981), and three of these elements are present in the gNa sequence, occurring at nucleotides 954-970, and 1004 (Fig. 4).

Comparison of the coding sequences of gNa and the cDNA clones cN1 and cN2 indicates that there are no introns and that all three coding sequences terminate with a single TAG codon. Within the coding sequence there is some divergence between the genomic and cDNA clones. For example, when the gNa sequence is aligned for maximum homology with the pN1 sequence it is observed that the genomic coding sequence is 21 nt longer than the cDNA. Excluding insertions, the two sequences are 90% homologous at the nucleotide level, with 17% of the nucleotide substitutions occurring in the third base of the codon. Alignment of the gNa- and pN2-derived peptide sequences shows 18 amino acids have been substituted excluding the gNa insertions but that only five of the substitutions are conservative (hydrophobic to hydrophobic, for example).

Expression of gNa.—Demonstrating the expression of a particular gene family member by hybridization requires a gene-specific probe. Since the nontranslated portions of genes often provide such probes, the 0.4 kb *XhoI*-*BamHI* fragment of gNa complementary to the 3' nontranslated portion gNa transcripts was nick-translated and used to probe duplicate genomic Southern blots of *EcoRI*-digested *B. napus* DNA (Fig. 5A, lanes 2 and 3). This probe hybridizes to just two napin genes at T_m -8°C (Fig. 5A, lane 2) and specifically to the 0.38 kb gNa *EcoRI* fragment at T_m -3°C (Fig. 5A, lane 3). Duplicate blots of size-fractionated *B. napus* embryo RNA were hybridized and washed in parallel with the DNA filters. Under the conditions that gave gene-specific DNA/DNA hybridization, a signal is detectable on the Northern blot corresponding to a napin-sized transcript (Fig. 5B, lane 2). No hybridization was evident when the DNA and RNA blots were washed at T_m -3°C, however (data not shown).

Mapping the 5' Terminus of the gNa Transcript.—Our first studies of the initiation site of gNa transcripts employed S1 nuclease digestion analysis (Fig. 6). The 0.38 kb *SalI*-*EcoRI* fragment of gNa was 5' end-labeled at the *SalI* site, and the labeled strand was purified on a polyacrylamide gel to use as a probe. The same fragment was sequenced to provide accurate electrophoretic size standards. Aliquots of this probe were hybridized at either T_m -25°C or T_m -4°C with 100 µg of *B. napus* embryo RNA. After digestion of the resulting hybrids the longest protected probe fragment was 138 bp, indicating

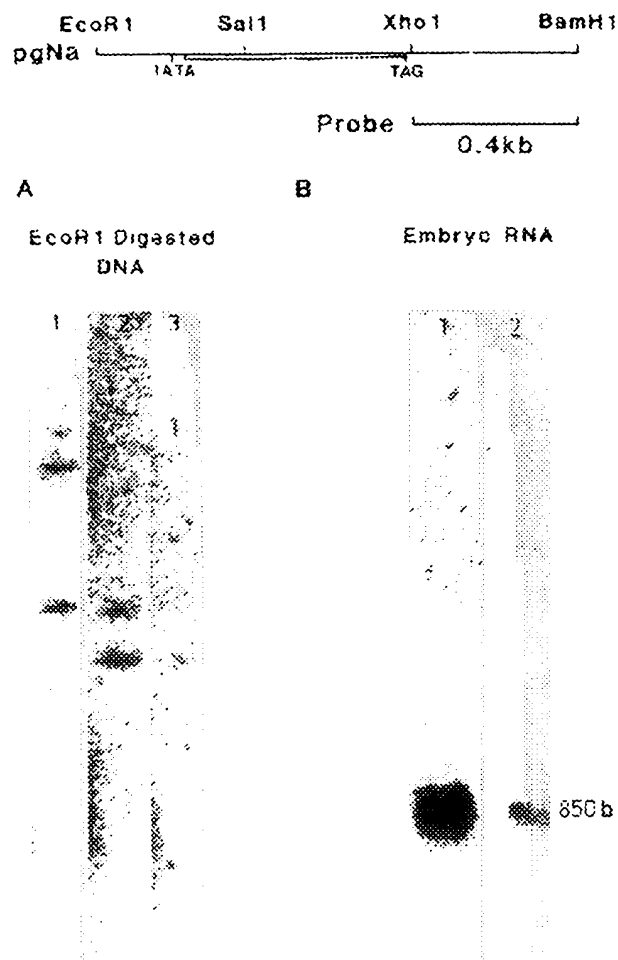


Fig. 7. Expression of gNa. A, lanes 1-3 are identical genomic Southern blots, each with 10 µg of *EcoRI*-digested *B. napus* DNA. Lane 1 was probed with nick-translated pN1 and washed at moderate stringency, 55°C in 0.1 × SSC, conditions that allow hybridization to the entire napin gene family. Lanes 2 and 3 show the gene-specific hybridization of the nick-translated 0.4-kb *XhoI*-*BamHI* fragment subcloned from gNa. Lane 2 was washed at T_m -3°C and lane 3 at 65°C in 0.1 × SSC. B, duplicate Northern blots with 0.25 µg of total *B. napus* embryo RNA probed and washed as in panel A, lanes 2 and 3. Under the same condition that give specific gene-specific hybridization in panel A, a napin size transcript is detected in panel B.

initiation at the T number 258 in Fig. 4. Also apparent are strong signals corresponding to cleavage in the 2 blocks of dA residues located 4 and 10 bp downstream from the initiation site, but the reason for cleavage at these sites is unclear. Local denaturation in the AT-rich regions seems unlikely as these signals are generated under nonstringent S₁ nuclease digestion conditions. It is possible that these signals represent other initiation points for the same gene or different 5' end structures of transcripts from other napin genes which are able to hybridize with the probe, which does contain 88 bp of coding sequence.

Primer extension analysis employing a synthetic oligonucleotide primer without coding sequence was undertaken to more specifically define the 5' end of the gNa transcript (Fig. 7). An oligomer was synthesized that was complementary to the 15 bases immediately 5' to the gNa codon. When hybridized to embryo RNA this primed the reverse transcription of a product extending 22 bases beyond the oligomer indicating RNA initiation at the dA numbered 259 in Fig. 4, just one

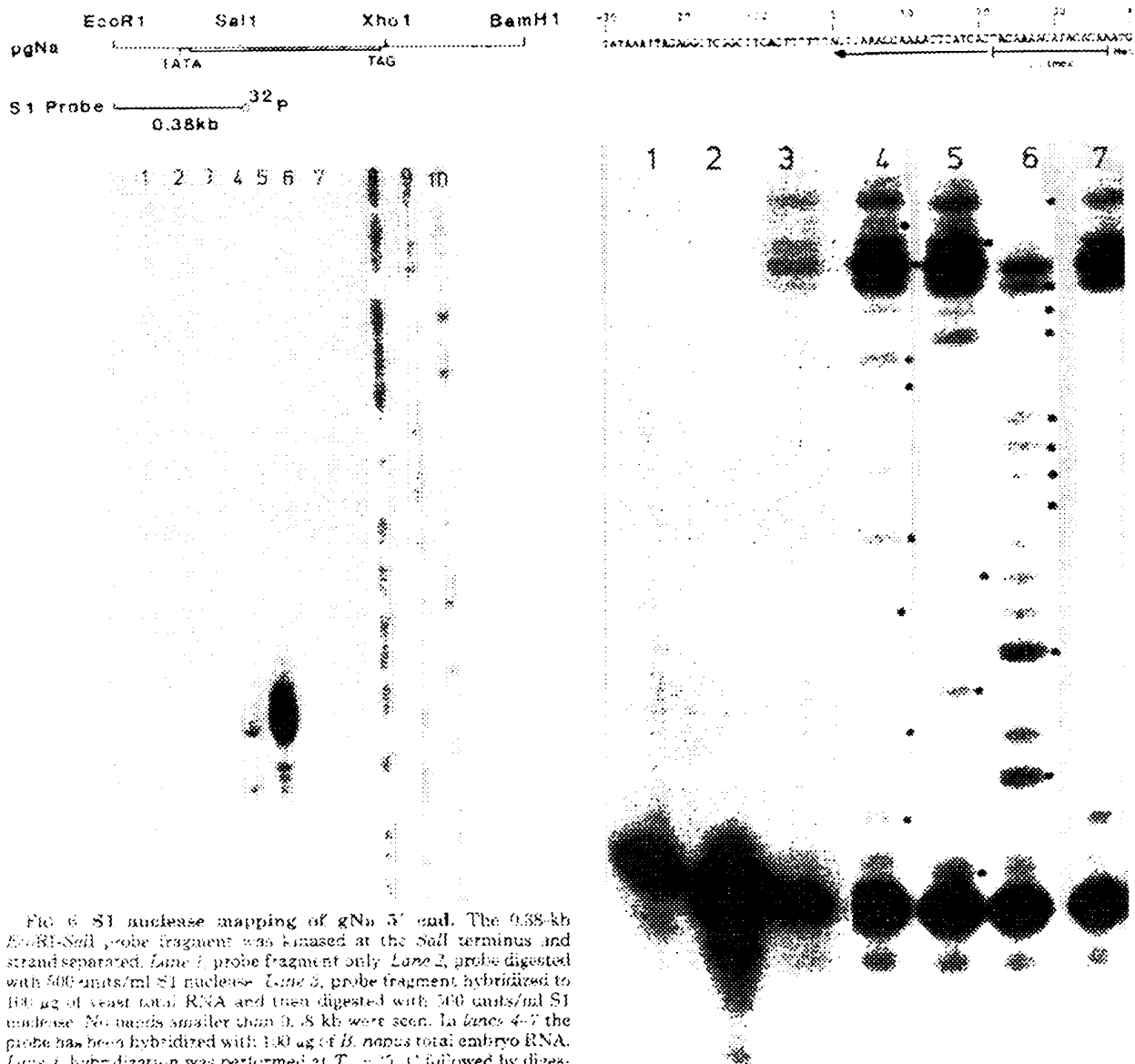


FIG. 6. S1 nuclease mapping of gNa 5' end. The 0.38-kb EcoR1-Sall probe fragment was kinased at the Sall terminus and strand separated. Lane 1, probe fragment only. Lane 2, probe digested with 500 units/ml S1 nuclease. Lane 3, probe fragment hybridized to 100 µg of yeast total RNA and then digested with 500 units/ml S1 nuclease. No bands smaller than 0.38 kb were seen. In lanes 4-7 the probe has been hybridized with 100 µg of *B. napus* total embryo RNA. Lane 4, hybridization was performed at $T_m - 25^\circ\text{C}$ followed by digestion with 500 units/ml S1 nuclease. Lane 5, hybridization was performed at $T_m - 25^\circ\text{C}$ and then digested with 100 units/ml S1 nuclease. Lane 6, hybridization was performed at $T_m - 8^\circ\text{C}$ and then digested with 500 units/ml S1 nuclease. Lane 7, the hybridization was performed at $T_m - 8^\circ\text{C}$ followed by digestion with 100 units/ml S1 nuclease. The 0.38-kb probe fragment was sequenced to provide size standards. Lane 8, AG reaction. Lane 9, TC reaction. Lane 10, G reaction.

base short of the 5' end and mapped by S1 nuclease protection. Since eucaryotic mRNAs are capped at purine residues (Cory and Adams, 1975), we expect the authentic RNA initiation site of gNa transcripts to be the dA₂₅₉ (Fig. 4), indicated by primer extension analysis. The gNa transcript thus has a 5' nontranslated leader 37 nucleotides long.

The primer extension experiments were also used to address the expression of gNa by performing the reverse transcription in the presence of dideoxynucleotides to determine the sequence of the primer extension product (Fig. 7). A sequence consistent with the expression of gNa can be detected, although the extent to which this portion of the gNa sequence is conserved among the napin genes is not yet known. The presence of heterogeneous signals in the sequence

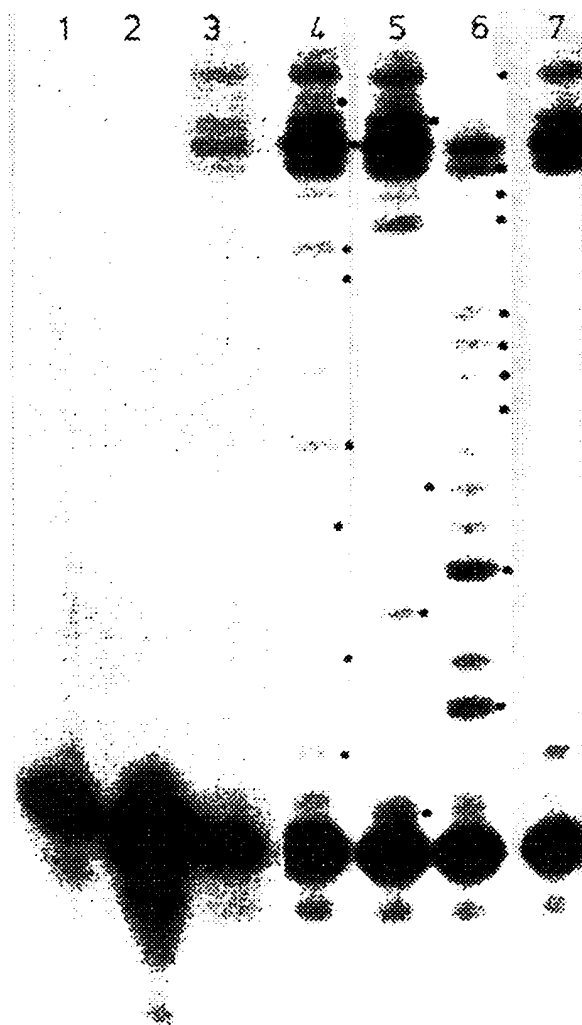


FIG. 7. Primer extension analysis of gNa. The nucleotide sequence upstream of the napin initiation codon is shown at the top of the figure. A 15-mer complementary to the gNa sequence 5' to the initiation codon was kinased and annealed to 100 µg of total embryo RNA. Hybrids were extended by avian myeloblastosis virus reverse transcriptase. Lane 1, primer only. Lane 2, primer extension with no RNA. Lane 3, primer extension using 10-fold less primer (0.1 ng) than lanes 4-7. Lanes 4, 5, 6, and 7, dideoxynucleotide sequencing of the extension product. Lane 4, G reaction. Lane 5, A reaction. Lane 6, T reaction. Lane 7, C reaction. Bands have been placed on the sequencing ladder where bands should occur if gNa transcripts were as template in this experiment.

ing ladder indicates that the primer hybridized with other napin transcripts as well.

DISCUSSION

Of the approximately 16 napin genes in *B. napus*, one has now been sequenced by us, gNa, and another, napA, by L.-G. Josera et al. In addition we previously reported the sequences

of L.-G. Josera in personal communication.

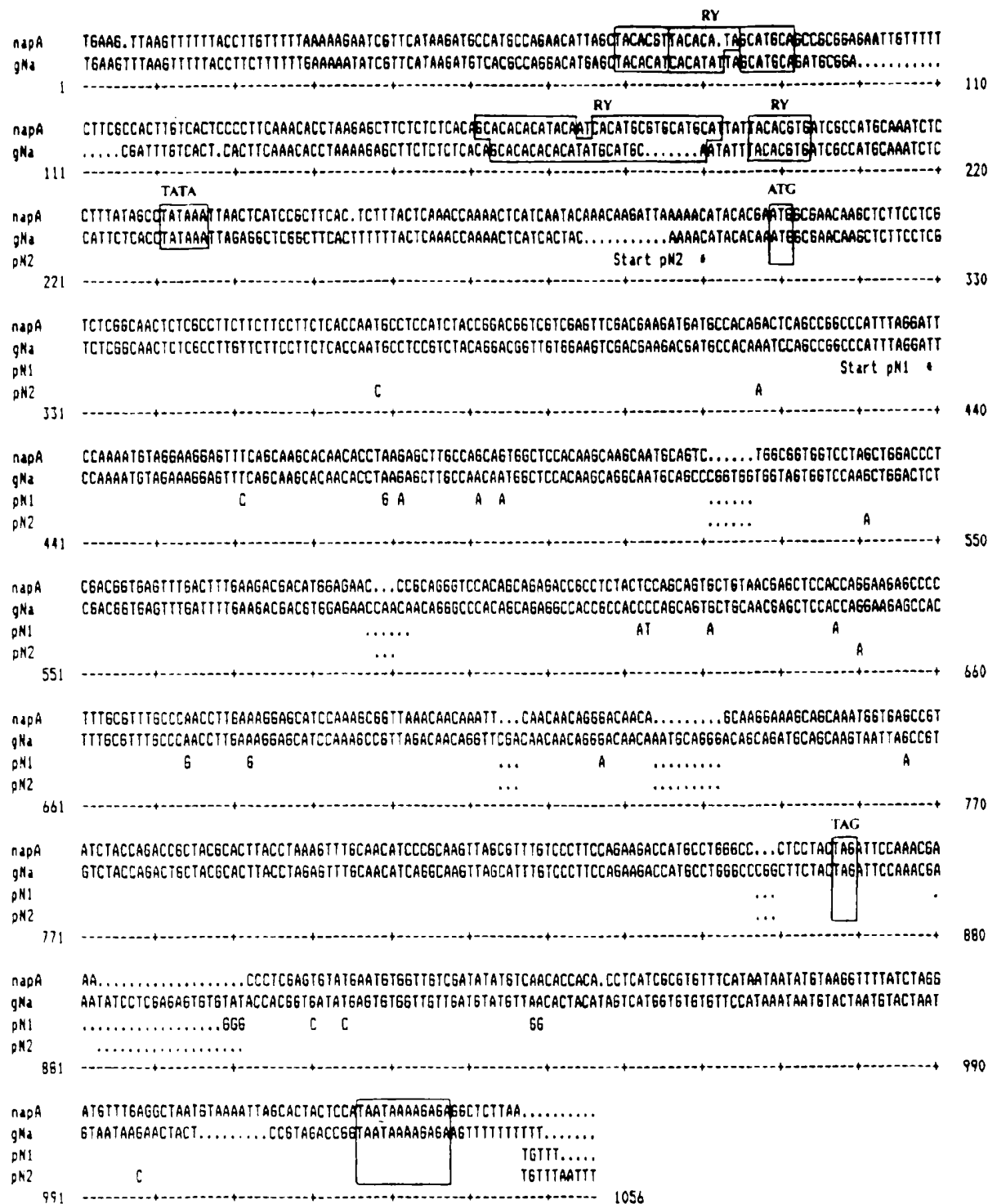


FIG. 8. One kb of the *gNa* genome sequence has been aligned for maximum homology with *napA* and two cDNA clones, pN1 and pN2. To emphasize the close homology between the cDNAs and *napA* only the cDNA bases that differ from *napA* have been displayed. Dots indicate positions where gaps have been introduced into a sequence for alignment purposes. Conserved features which have been designated are: the alternating purine-pyrimidine blocks (RY), the TATA boxes, the initiation and termination codons, and the 12 bp of homology shared at the most downstream consensus sequence associated with polyadenylation.

of two different cDNA clones representing transcripts from other genes (Crouch *et al.*, 1983). Thus, four members of the family have been examined, although their relative levels of expression are not known. Comparison of all four coding sequences (Fig. 8) indicates that the cDNAs and napA are greater than 95% homologous. The gNa sequence with insertions at positions 521, 588, 714, and 734 of Fig. 8 is likely to represent a minor class of napins, perhaps one of the four discrete species fractionated by Lonnerdal and Janson (1972).

The 3' nontranslated regions of the cDNAs and napA are as highly conserved as the coding regions. Such high homology would preclude the use of these sequences for gene-specific hybridization as was possible for gNa. One of the distinctive features of this portion of gNa is the presence of three sequences resembling the consensus associated with polyadenylation of mRNAs. It is striking that although the gNa 3' nontranslated region is divergent, all four napin sequences are perfectly homologous for 12 bp around the most downstream consensus polyadenylation element, suggesting that this is the authentic polyadenylation signal for the genomic clones.

The nucleotide sequence of the genomic clone gNa and its flanking regions contain the canonical features expected of plant genes transcribed by RNA polymerase II (Messing *et al.*, 1983). There are no introns, which is characteristic of genes for many of the other 2 S seed proteins and related cereal prolamins. In the 5' flanking region of gNa are several blocks of alternating purine-pyrimidine nucleotides, which have been observed in viral enhancer (Lusky *et al.*, 1983). Their significance in napin genes remains to be tested.

Alignment of the two napin genomic clones for maximum homology (Fig. 8) shows that the coding sequence of gNa is 24 bp longer than napA with the extra sequence occurring as three additions of single codons and two insertions of two codons. However, the 5' RNA leader region of gNa is deleted by 10 nucleotides relative to napA. As already mentioned, the two genomic sequences diverge sharply past the coding sequence termination codons. In contrast, the 5' flanking region is highly conserved overall, including the regions of alternating purine-pyrimidine residues. Since the entire 5' flanking region is so highly conserved, it is difficult to single out regions by comparative homology that might be involved in the temporal or spatial regulation of napins.

As mentioned earlier, napin is evolutionarily related to some of the cereal prolamin storage proteins. However, there is no evidence in napin genomic sequences of homology to the short upstream sequences found to be conserved in the genes for α -gliadin, β -hordein, and the (unrelated) zeins (Forde *et al.*, 1985). If the conserved prolamin sequence is functionally significant, its absence in napin may be related to the difference in spatial expression; napin is synthesized in the embryo, whereas prolamins are restricted to endosperm cells.

Acknowledgments—We wish to thank Dr. Lars-Göran Josefsson, Wallenberg Laboratory, Uppsala, Sweden, for exchanging napin nu-

cleotide sequences prior to publication. We appreciate the synthesis of oligonucleotides by Lawrence Washington, Institute for Cell and Molecular Biology, with funds from Indiana Corporation for Science and Technology, and the expert advice of Karen Tenbarger, Jerome Cane, and Lorraine Solberg. Most helpful discussions were provided by Drs. Keith Blundy and Vic Knauf, Calgene Inc., Davies, CA. We are grateful for Karen Parr's swift assistance in the preparation of this manuscript.

REFERENCES

- Benoist, C., O'Hare, K., Breathnach, R., and Chambon, P. (1980) *Nucleic Acids Res.* **8**, 127-142.
- Benton, W. D., and Davis, R. W. (1977) *Science* **196**, 180-184.
- Bruskin, A. M., Tyner, A. L., Wells, D. E., Showman, R. M., and Klein, W. H. (1981) *Dev. Biol.* **87**, 308-318.
- Cory, S., and Adams, J. M. (1975) *J. Mol. Biol.* **99**, 519-547.
- Crouch, M. L., Tenbarger, K. M., Simon, A. E., and Ferl, R. (1983) *J. Mol. Appl. Genet.* **2**, 273-283.
- Crouch, M. L., Tenbarger, K., Simon, A., Finkelstein, R., Scotfield, S., and Solberg, L. (1985) *Molecular Form and Function of the Plant Genome* (Van Vloten-Doting L., ed) pp. 555-566, Plenum Publishing Corp., New York.
- Dove, W. F., and Davidson, N. (1962) *J. Mol. Biol.* **5**, 467-478.
- Favaloro, J., Treisman, R., and Kamen, R. (1980) *Methods Enzymol.* **65**, 718-749.
- Finkelstein, R. R., Tenbarger, K. M., Shumway, J. E., and Crouch, M. L. (1985) *Plant Physiol.* **78**, 630-636.
- Fitzgerald, M., and Shenk, T. (1981) *Cell* **24**, 251-260.
- Forde, B. G., Heyworth, A., Pywell, J., and Kreis, M. (1985) *Nucleic Acids Res.* **13**, 7327-7339.
- Frischauf, A.-M., Lehrach, H., Poustka, A., and Murray, N. (1983) *J. Mol. Biol.* **170**, 827-842.
- Higgins, T. J. V. (1984) *Annu. Rev. Plant Physiol.* **35**, 191-221.
- Jay, E., Seth, A. K., Rommens, J., Sood, A., and Jay, G. (1982) *Nucleic Acids Res.* **10**, 6319-6329.
- Karn, J. M., Brenner, S., Barnett, L., and Cesareni, G. (1980) *Proc. Natl. Acad. Sci. U. S. A.* **77**, 5172-5176.
- Kasadara, D. D., Okita, T. W., Bernardin, J. E., Baelker, P. A., Nimmo, C. C., Lew, E. J., Dietler, D. D., and Green, F. C. (1984) *Proc. Natl. Acad. Sci. U. S. A.* **81**, 4712-4716.
- Kreis, M., Forde, B. G., Rahman, S., Miflin, B. J., and Shewry, P. R. (1985) *J. Mol. Biol.* **183**, 499-502.
- Lonnerdal, B., and Janson, J.-C. (1972) *Biochim. Biophys. Acta* **278**, 175-183.
- Lusky, M., Berg, L., Weiher, H., and Botchan, M. (1983) *Mol. Cell. Biol.* **3**, 1108-1122.
- Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning, A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
- Maxam, A. M., and Gilbert, W. (1980) *Methods Enzymol.* **65**, 499-560.
- Messing, J., Geraghty, D., Heidecker, G., Hu, N.-T., Kridl, J., and Rubenstein, I. (1983) in *Genetic Engineering of Plants* (Meredith, C. P., Kosuge, T., and Hollaender, A., eds) pp. 211-226, Plenum Publishing Corp., New York.
- Odani, S., Koide, T., and Ono, T. (1983) *Biochem. J.* **213**, 543-545.
- Scalenghe, F., Turco, E., Edstrom, J. E., Pirrotta, V., and Melli, M. (1981) *Chromosoma* **82**, 205-216.
- Sharief, F. S., and Li, S. S.-L. (1982) *J. Biol. Chem.* **257**, 14753-14759.
- Verma, S. C., and Rees, H. (1974) *Heredity* **33**, 61-68.
- Vieira, J., and Messing, J. (1982) *Gene (Amst.)* **19**, 259-268.
- Youle, R. J., and Huang, A. H. C. (1981) *Am. J. Bot.* **68**, 44-48.

Continued on next page.

Supplementary Material to:
NUCLEOTIDE SEQUENCE OF A MEMBER OF THE NAPIN STORAGE PROTEIN FAMILY FROM *BRASSICA NAPUS*
Steven R. Scofield and Martha L. Uebersch.

MATERIALS AND METHODS

Plants

Brassica napus L. cv. Tower seeds (from Dr. W.D. Baverstock, University of Guelph, Ontario) were planted in a 2:1:1 (by volume) mixture of soil, vermiculite and perlite. The plants were grown under greenhouse conditions.

Unexpanded first leaves of seedlings were harvested and frozen under liquid nitrogen. Forty g of leaves were ground by mortar and pestle, then homogenized for 1 minute at high speed in a Waring blender, again under liquid nitrogen. The resulting powder was suspended in 2.0 l M Tris HCl pH 8.5, 3.0 l M EDTA, 0.08 M KCl, 3.5 M sucrose, 0.004 M spermidine, 3.00 l M spermine, 0.30 l M phenylmethylsulfonyl fluoride, 0.05 l 2-mercaptoethanol and 0.25 l Tris HCl pH 8.0. The suspension was filtered through 40 µm mesh nylon filter cloth (Hytest), and the resulting filtrate was centrifuged at 2000 X G in a Sorvall NMR rotor. Three cycles of differential sedimentation centrifugation were typically performed to purify nuclei. The pelleted nuclei were resuspended in 25 ml of nuclear isolation buffer and then lysed by the addition of 25 ml of 1% Sarkosyl followed by the immediate addition of 0.97 g/l CaCl₂. Polysaccharides were removed from the solution by centrifugation at 13,000 X G in a Sorvall SS34 rotor. The supernatant was collected, sodium bromide (NaBr) was added to a final concentration of 10 µg/ml and the refractive index was adjusted to 1.395. After 2 rounds of equilibrium centrifugation at 40,000 rpm for 40 h in a Beckman T150 rotor, NaBr was extracted from the banded DNA with 1-butanol. The DNA was then precipitated by 3 volumes of 70% EtOH at -20°C.

Isolation of RNA

Total RNA from *B. napus* embryos, 15-10 days post-anthesis, was prepared by extraction with phenol and precipitation with lithium chloride, as described in detail in Pinalis et al. (1985).

Genomic DNA hybridization

Genomic Southern blot analysis was typically performed using 10 µg of DNA per gel lane. Restriction experiments were based on 10 - 1.5 µg nuclear DNA (Varma and Bawa, 1976). Genomic DNA was digested with 3 units of restriction enzyme per µg DNA for at least 3 h, and completeness of digestion was monitored by including 1 µg of phage λ DNA in the reactions. If the pattern expected for completely digested λ DNA was seen superimposed on the genomic DNA pattern it was assumed that unaltered digestion had occurred. The restricted DNA was electrophoresed on 0.4 cm thick, 2.8% agarose gels in 0.04 M Tris acetate 0.20 M EDTA (TAE) buffer (Maniatis, 1982) for 12 h at 30 V, and then transferred to nitrocellulose according to Southern (1975). Filters were prehybridized for at least 4 h in 7x SSC, 5x Denhardt's (1x Denhardt's is 0.02% each Pirlol, bovine serum albumin and polyvinylpyrrolidone), 50% deionized formamide, 0.1 M sodium hydrogen phosphate, 500 µg/ml sheared calf thymus DNA and 25 µg/ml polyribonucleic acid at 37°C. Hybridization was performed in the above solution except for including 10% dextran sulfate (5000 MW) and reducing calf thymus DNA to 100 µg/ml. Fifty µl of hybridization solution were used for each cm² of filter and 0.5 to 1.0 µg of nick-translated probe, specific activity 3.2-10⁷ cpm/µg, were typically added to the hybridization. The filters were hybridized for 12-20 h at 37°C, and washing was according to Maniatis (1982). The stringency was controlled by varying the final wash temperature. It was calculated by the equations of Davie and Davidson (1982); the GC content of pN1 to 60%.

RNA hybridization

Total *B. napus* embryo RNA, 0.25 µg per gel lane, was electrophoresed and transferred to nitrocellulose as in Brinkley et al. (1981). The filters were included in the same hybridization reactions and washed as the genomic Southern filters described above.

Genomic libraries

Randomized DNA libraries representing *B. napus* 50 kb partial digestion fragments 10-22 kb in size were constructed in the phage λ vector BDGA (Fritsch et al., 1983) following to protocols of Maniatis (1982). *In vitro* packaging was carried out in extracts prepared as in Szallasi et al. (1981), the packaged phage were plated on Q59 (Kern et al., 1980). The library, consisting of 4x10⁷ recombinants, was screened with nick-translated pN1 probe by the method of Benton and Davis (1977). Subclones were constructed in pUC8 or pUC18 (Yi and Haining, 1982).

DNA sequencing

The sequence of gN was determined by the base-specific chemical cleavage method of Maxam and Gilbert (1980) with the modification of Jay et al. (1982). Twenty µl of 10% (vol/vol) acetylation were added to the first resuspension of all hydrazine reactions and incubated at room temperature for 5 min before proceeding with the second strand precipitation. This step was completely removed residual hydrazine which can cause cleavage at guanine bases during subsequent piperidine strand selection reactions.

5' nuclease mapping

The 3.38 kb *Sal*I-EcoRI restriction fragment of pN1 was ligated at the *Sal*I terminus (Maxam and Gilbert, 1980). This fragment was strand-separated on a 5% acrylamide, 0.1% bisacrylamide gel (Maniatis, 1982), and electrophoresed from a gel slice into a dialysis bag containing 1 M 0.09 M Tris-borate pH 8.0 and 0.002 M EDTA and 10 µg/ml BSA. The eluate was ethanol-precipitated and resuspended in distilled water for use as the 5' probe.

Approximately 10⁵ cpm of probe was mixed with 100 µg of *B. napus* total embryo RNA, ethanol-precipitated, washed with 70% ethanol and resuspended in 30 µl of hybridization buffer (0.04 M PIPES pH 6.4, 0.4 M NaCl, 0.001 M EDTA) (Famulski et al., 1980). The samples were heated to 75°C for 15 min and then transferred to water baths at either 46°C (70-75°C) or 62°C (70-6°C). After hybridizing for 12 h, 300 µl of 81 nucleosome eluate was added (0.28 M NaCl, 0.05 M Na acetate, 0.005 M ZnSO₄, 20 µg/ml denatured calf thymus DNA and 100 or 500 units/ml S1 nuclease). The samples were rapidly transferred to a 37°C water bath, incubated for 1 h, and then extracted with 1:1 phenol/chloroform and precipitated with isopropanol. The samples were analyzed on 6% polyacrylamide/urea sequencing gels.

Primer extension

An oligonucleotide with the sequence 5'-TTCCTACCTCTTCT-3' was synthesized on an Applied Biosystems RNA synthesizer operated by the Indiana University Institute of Molecular and Cellular Biology. The oligomer was 5' end labeled with gamma 32P ATP by T4 polynucleotide kinase treatment. Three µg of the labeled primer was mixed with 100 µg of *B. napus* embryo RNA in 30 µl of primer extension hybridization buffer (0.1 M Tris HCl pH 8.0, 0.01 M MgCl₂), heated to 10°C for 1 minute, then slowly cooled to 40°C, and ethanol-precipitated. The primer-RNA hybrids were resuspended for reverse transcription in a 25 µl reaction consisting of 0.1 M Tris HCl pH 8.3, 0.140 M KCl, 0.01 M MgCl₂, 0.5 mM dNTPs and 40 units of AMV reverse transcriptase, incubated at 41°C for 10 minutes, and then dried in a vacuum centrifuge (Speed-vac). The samples were then analyzed on 15% acrylamide/urea gels.

Dideoxynucleotide chain termination sequencing of the primer extension product was performed by modifying the reverse transcription reactions. Four 10 µl reverse transcription reactions were performed, one for each dNTP. They were as described above except dNTPs were 0.1 mM and dNTPs were added (only one to each reaction) at a concentration of 2.5 µM dNTP. After incubating 10 minutes at 41°C the reactions were changed for 15 minutes by addition of 2 µl of 1 mM dNTP.

A. Genomic Napin Clones

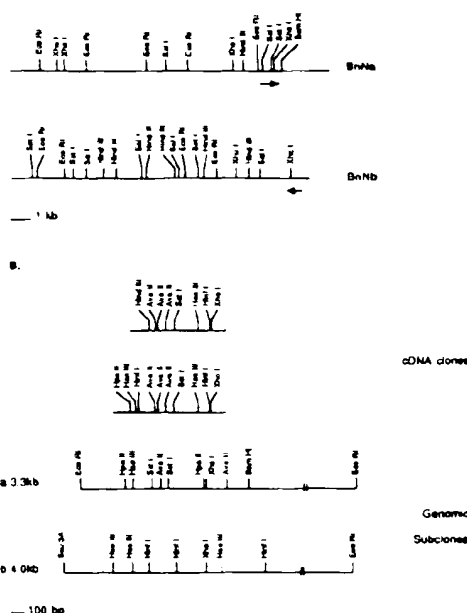


Figure 2. (A) Restriction maps of two *B. napus* DNA fragments isolated from genomic libraries that encode napin. Arrows indicate direction of napin transcription. (B) Comparison of restriction maps of cDNA clones pN1 and pN2 with genomic subclones pN1a and pN1b.

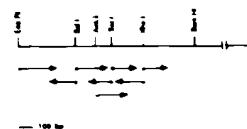


Figure 3. Sequencing strategy for gN. Open circles denote that both strands were sequenced from that restriction site. In separate reactions these sites were 5' labeled with T4 polynucleotide kinase or 3' labeled with AMV reverse transcriptase. At closed circles only 5' ends were labeled.

20. Bentley, R. (1983) *Proc. Natl. Acad. Sci. USA* 80, 1416-1420.
21. Bentley, R. and Jambick, B. (1983) *Methods Enzymol.* 109, 409-433.
22. Bentley, R. and Jambick, B. (1984) *Proc. Natl. Acad. Sci. USA* 81, 1072-1076.
23. Bentley, R. (1983) *Cell* 32, 111-119.
24. Bentley, R., Jambick, B., Jambick, B., and Jambick, B. (1983) *Proc. Natl. Acad. Sci. USA* 80, 1421-1425.
25. Bentley, R., Jambick, B., Jambick, B., and Jambick, B. (1983) *Proc. Natl. Acad. Sci. USA* 80, 1426-1430.
26. Bentley, R. (1983) *In* *Methods of Enzymology and Biochemistry*, 109, 409-433.
27. Bentley, R., Jambick, B., Jambick, B., and Jambick, B. (1983) *Proc. Natl. Acad. Sci. USA* 80, 1431-1435.
28. Bentley, R., Jambick, B., Jambick, B., and Jambick, B. (1983) *Proc. Natl. Acad. Sci. USA* 80, 1436-1440.
29. Bentley, R., Jambick, B., Jambick, B., and Jambick, B. (1983) *Proc. Natl. Acad. Sci. USA* 80, 1441-1445.
30. Bentley, R., Jambick, B., Jambick, B., and Jambick, B. (1983) *Proc. Natl. Acad. Sci. USA* 80, 1446-1450.

Nucleotide sequence of a B1 hordein gene and the identification of possible upstream regulatory elements in endosperm storage protein genes from barley, wheat and maize

B.G. Hooley, A. Heyworth, J. Powell and M. Kreis

Biochemistry Department, Rothamsted Experimental Station, Harpenden, Herts AL5 2JQ, UK

Received 19 July 1985; Revised and Accepted 30 September 1985

ABSTRACT

The B-hordeins are the major group of prolamins storage proteins in barley (*Hordeum vulgare* L.) and they are encoded by a small multigene family that is expressed specifically in the developing endosperm. We report the complete nucleotide sequence of a clone of one B-hordein gene (pBHR184). The cloned gene contains no introns and belongs to the B1 sub-family of B-hordein genes. Comparison of the 5'-flanking sequences of pBHR184 with those of related 5-rich prolamins genes from wheat shows that several short sequences within 600 bp upstream of the translation initiation codon are strongly conserved. A sequence that is conserved at around -300 bp in the 5-rich prolamins is also conserved at similar locations in genes encoding the two major classes of maize prolamins (the Z19 and Z21 zeins) and appears to be unique to prolamins genes. We discuss the possible role of this '-300 element' in the control of gene expression in the developing cereal endosperm.

INTRODUCTION

In most cereal species the major seed storage proteins are prolamins, a complex group of alcohol-soluble polypeptides that make up about half of the protein in the mature grain. In barley (*Hordeum vulgare* L.) they are classified into three main groups (B-, C- and D-hordeins), which are specified by separate compound genetic loci on chromosome 5 (1). Prolamins homologous to the hordeins are found in wheat (the gliadins and glutenins) and in rye (the secalins), while the major prolamins of the more distantly related maize (the zeins) seem to have evolved independently (1,2).

Synthesis of the prolamins polypeptides is endosperm-specific and is initiated coordinately at a relatively late stage of seed development (3-5). In barley, expression of some families and sub-families of hordein genes is modulated by the balance of nitrogen and sulphur nutrition (6) and by a mutation at an unlinked 'regulatory' locus (7,8). In maize, mutants that alter either the timing (9) or the rate (10) of zein deposition have been reported, some of which specifically affect synthesis of one or other of the two zein classes. Thus it appears that there are at least two types

of control operating on prolamin gene expression, one responsible for coordinate induction of the genes during endosperm development and another regulating the subsequent rate of prolamin accumulation, and these controls have the ability to act differentially on subsets of prolamin genes.

As part of our study of the organization and expression of the hordein gene families we now report the isolation and nucleotide sequencing of a B-hordein genomic clone. We discuss the possible significance of short upstream sequences that are conserved in B-hordein and α -gliadin genes and in genes encoding the two major classes of zeins.

METHODS

Screening a barley genomic library

A genomic library of barley DNA (*Hordeum vulgare* L., cv. Sundance) was generously provided by Dr. M. Murray and Dr. J. Slightom (Agrigenetics Corporation, Madison). The unamplified library (1×10^6 recombinant phage), which had been prepared by cloning a partial *EcoRI* digest of high molecular weight barley DNA in Charon 32, was screened by plaque hybridization (11) using as probe the nick-translated (12) cDNA inserts from pB7 and pB11 (13). Hybridizing clones were plaque-purified and phage DNA was prepared by a plate-lysate method (14).

Nucleotide sequencing

A 2.9 kb *EcoRI* fragment from λ HVBH3.4 was sub-cloned in pUC9 for sequencing. Plasmid DNA of the subclone (pBHR184) was prepared from cells lysed with Triton X-100 (15) and further purified by banding twice in CsCl gradients. Fragments suitable for sequencing were generated by BAL-31 deletion. Three μ g of pBHR184 was linearized with the appropriate restriction enzyme (see Fig. 1) and digested at a rate of 80 bp/min/end using 0.7 U BAL-31 at 37° in a 60 μ l volume (16). Aliquots taken at 2 min intervals up to 18 min were phenol-extracted, digested with a second restriction enzyme and cloned in *E. coli* strain JM101 using as vectors M13 mp8 (17) or mp19 (supplied by Pharmacia Ltd.). Sequencing was by the dideoxy method (18) and sequences were assembled and analysed with the assistance Staden programs (19-21) operating on a VAX 11/750 computer.

SI mapping

Fragments of the cloned B hordein gene were prepared for SI protection analysis as follows. Single-stranded phage DNA was prepared from two M13 clones that contained the 2.9 kb *EcoRI* fragment from pBHR184 in opposite orientations. Three μ g of each phage was mixed in 10 μ l of 50 mM Tris HCl

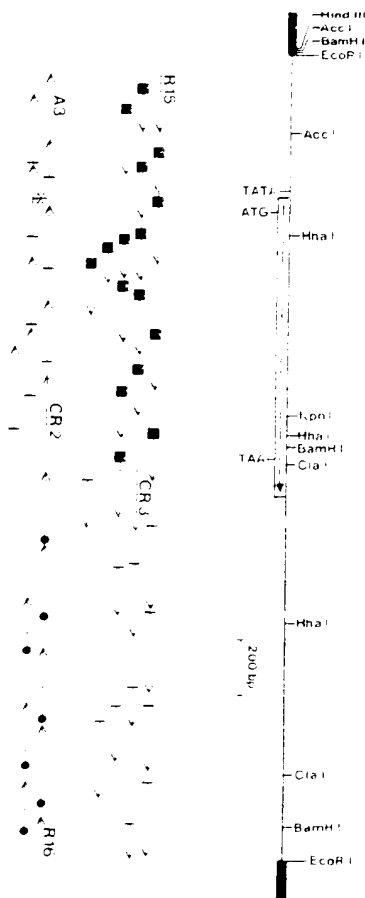
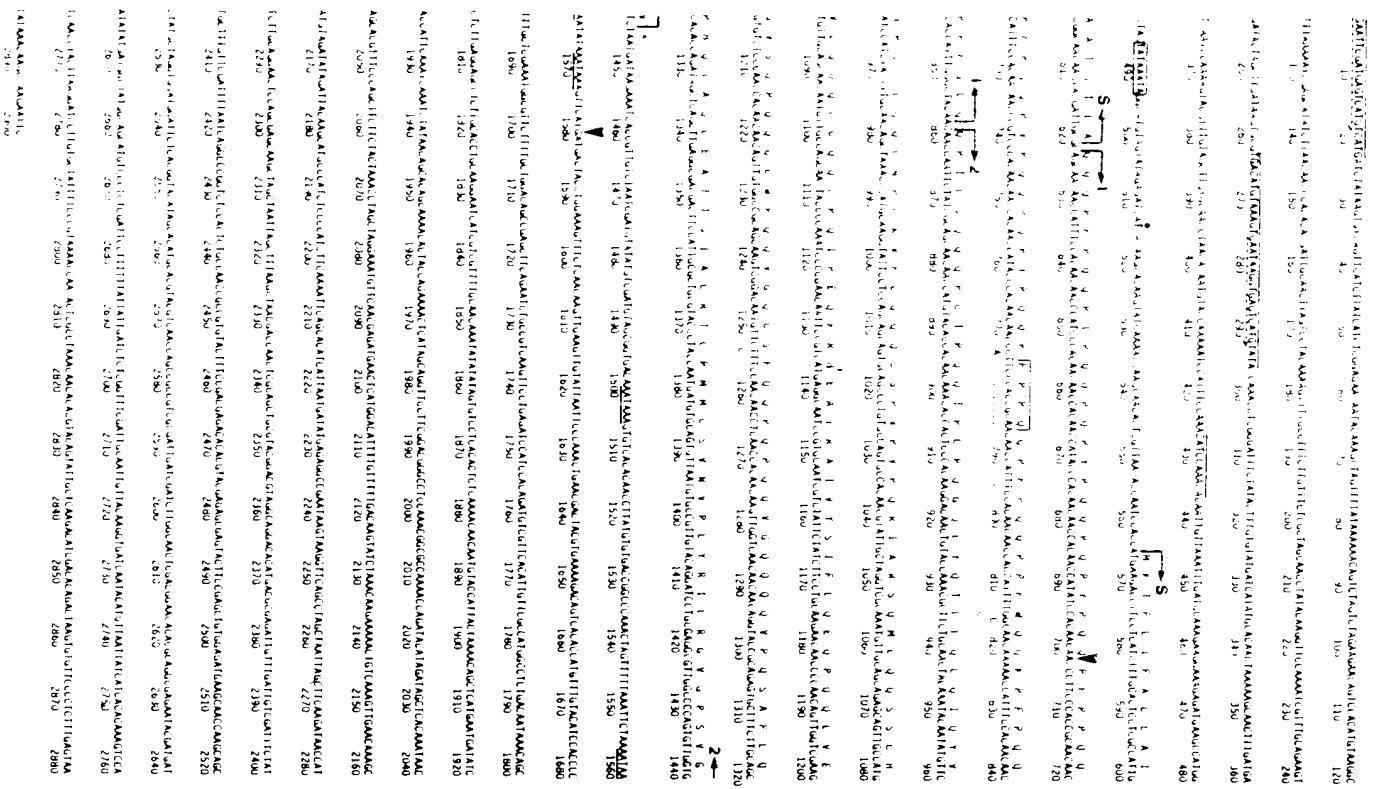


Fig. 1. Restriction map of a B-hordein genomic clone (pBHR184) and the sequencing strategy. The clone was constructed by sub-cloning a 2.9 kb *EcoRI* fragment from λ HVBH3.4 into pUC9. The positions of the *AccI* and *BamHI* sites were determined experimentally and the locations of the other sites were obtained from the sequence. The open rectangle indicates the region corresponding to the mature mRNA (as deduced from subsequent analysis) and the arrow within it shows the direction of transcription. Arrows with closed squares indicate sequences obtained from M13 sub-clones generated by digestion of pBHR184 with, in turn, *HindIII*, *BAL-31* and *EcoRI* (see Methods). For sequences indicated by arrows with closed circles the order was: *EcoRI*, *BAL-31*, *ClaI*; for those indicated by arrows with vertical bars it was: *ClaI*, *BAL-31*, *EcoRI*. The remaining sequences were obtained by sub-cloning restriction fragments without *BAL-31* digestion. The asterisk indicates the M13 sub-clone (*ClaI*42) that was used in Fig. 3 to provide size markers.

pH8.0, 50 mM NaCl and incubated at 60° for 30 min. The mixture was made to 10 mM MgCl₂ and the DNA digested with 4 U *HhaI*. Because the annealed phage molecules are only double-stranded for the length of their inserts the *HhaI* digest generates only three fragments. The *HhaI* fragments (3 μ g) were end-labelled with γ -³²P using T4 polynucleotide kinase (Bethesda Research Laboratories Inc.). Hybridization to 7 μ g poly A⁺ RNA from barley endosperm (cv. Sundance), and subsequent treatment with SI nuclease, were according to Berk and Sharp (22). Size markers were generated by performing sequencing reactions on an M13 sub-clone of pBHR184 (*ClaI*42) that contains the 5'-flanking region of the gene and a short region of coding sequence, including the *HhaI* site immediately downstream from the translation initiation codon (see Fig. 1). To generate dideoxy-terminated fragments with the same 5' end as the protected fragment to be mapped, the products of the sequencing reaction were treated with *HhaI* and *BamHI* before electrophoresis. *BamHI* cuts within the polynucleotide region of *ClaI*42 and was used to reduce the size of the fragments that contained the sequencing



primers, some of which would otherwise have co-migrated with the size-markers.

RESULTS AND DISCUSSION

Cloning and sequence analysis of a B1 hordein gene

A barley genomic library was screened for *b*-hordein genes using a mixed probe consisting of two cDNA clones, pB7 and pB11, which represent the two major sub-families of B-hordein mRNA (13). Of three clones that were plaque-purified and mapped by restriction digests, one was selected for detailed analysis. This clone, *3HvBH3.4*, contains a barley DNA fragment of 17.4 kb, which is made up of four *EcoRI* fragments of 7.4 kb, 3.6 kb, 3.5 kb and 2.9 kb. Southern blots (23) of the *EcoRI* digest revealed that only the latter fragment hybridized to the cDNA probe (not shown). Fig. 1 shows a restriction map of a sub-clone of the 2.9 kb fragment (pBHR184) and the sequencing strategy.

The nucleotide sequence of the 2.9 kb fragment is presented in Fig. 2. Nucleotides 564 to 1442 constitute an open reading frame that begins with an AUG codon and encodes a B-hordein-like polypeptide. Comparison with the sequences of the two cDNA clones that were used as probes shows that pBHRI84 is much more closely related to pB11, a B1-type hordein cDNA clone, than to pB3, a B3-type (13). The alignment between pBHRI84 and pB11 (which is incomplete at the 5' end) extends from nucleotides 701 to 1579, where the poly(A) tail in the cDNA clone begins. The coding sequence of the genomic fragment contains a 12 nucleotide sequence (positions 774-785) that is absent in pB11. There are only 4 other mismatches between the two sequences, and only one of these is a replacement substitution (Fig. 2). The differences between the cDNA and genomic sequences are consistent with previous evidence indicating nucleotide and amino acid sequence variations in the B-hordein multigene family (12, 24). As with other cereal prolamins

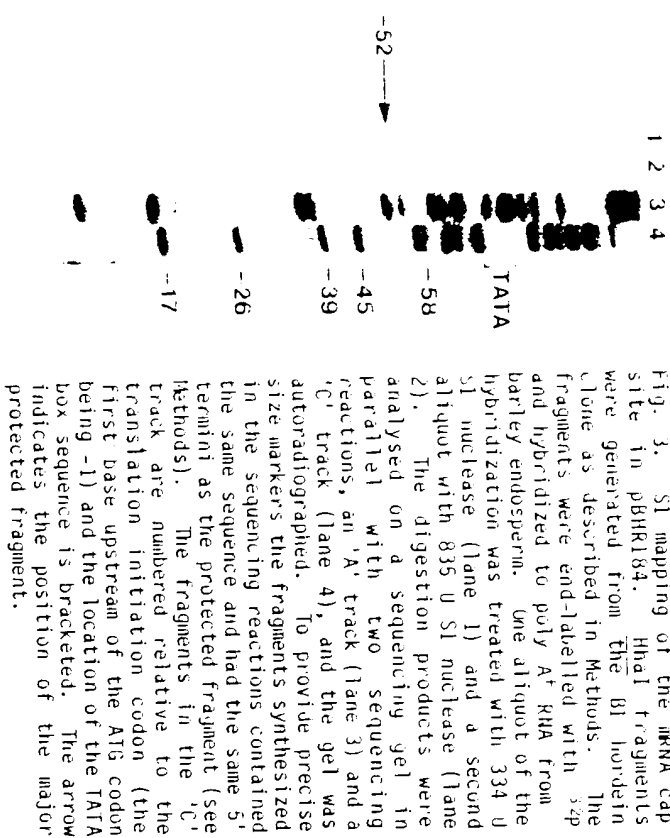


Fig. 3. SI mapping of the mRNA cap site in pBHR184. *Hha*I fragments were generated from the B1 hordein clone as described in Methods. The fragments were end-labelled with 32 P and hybridized to poly A⁺ RNA from barley endosperm. One aliquot of the hybridization was treated with 334 U *S*I nuclease (lane 1) and a second aliquot with 835 U *S*I nuclease (lane 2). The digestion products were analysed on a sequencing gel in parallel with two sequencing reactions, an 'A' track (lane 3) and a 'C' track (lane 4), and the gel was autoradiographed. To provide precise size markers the fragments synthesized in the sequencing reactions contained the same sequence and had the same 5' termini as the protected fragment (see Methods). The fragments in the 'C' track are numbered relative to the translation initiation codon (the first base upstream of the ATG codon being -1) and the location of the TATA box sequence is bracketed. The arrow indicates the position of the major protected fragment.

genes, there is no evidence for the presence of introns (25-32).

In the 3'-untranslated region of the gene there are several hexanucleotide sequences (AAATAA) that conform to the putative polyadenylation signal sequence (33) and there may therefore be several alternative polyadenylation sites. By analogy with pB11, which is identical to pBHR184 in the 3'-untranslated region, it is likely that there is a polyadenylation site at position 1579.

We have used an *S*I protection assay to map the mRNA cap site in pBHR184. A *Hha*I digest of the cloned gene was end-labelled with 32 P and annealed to poly A⁺ RNA from barley endosperm under R-looping conditions. The fragments were then treated with two concentrations of *S*I nuclease and analysed on a sequencing gel (Fig. 3, lanes 1 and 2). Precise mapping of the cap site was achieved by using size markers that contained the same sequence as the protected fragments (Fig. 3, lanes 3 and 4). The major protected fragment migrates at the position corresponding to 52 bp upstream from the translation initiation codon. Allowing one base for the mRNA cap, we estimate the transcription start site to be at position -51 relative to the ATG codon (see Fig. 2). Assuming a poly(A) tail of 80 residues (34),

the predicted length of the mRNA is 1150 nucleotides, which is in good agreement with previous estimates based on Northern blots (35).

Primary structure of the pBHR184 gene product

The open reading frame that starts at nucleotide 564 (Fig. 2) encodes a protein with 293 residues (M_r 33,423). The amino terminal region of the protein has many of the characteristics of a signal peptide, including a charged residue near the N-terminus and a core of hydrophobic residues (36). The presence of a signal peptide would be consistent with the evidence that the B-hordeins are synthesized on the rough endoplasmic reticulum and deposited in protein bodies (34,37,38), and with the observations that they are synthesized *in vitro* as larger precursors (34,39). However, because the N-termini of the B-hordeins are blocked (40,41), the N-terminal sequence of the mature protein is not known and we cannot assign the site of signal peptide cleavage with certainty. Nevertheless, comparison with the nascent and mature sequences of the homologous α -gliadin storage proteins from wheat (42) suggests that cleavage would occur between residues 19 and 20. The same cleavage site is predicted by application of the rules formulated by von Heijne (43). The mature protein would therefore consist of 274 residues and have a M_r of 31,444, which agrees well with estimates based on direct analysis of the B-hordeins (40).

Immediately following the putative signal peptide there is a region of the protein that is extremely rich in proline and glutamine. This part of the protein was previously identified as domain 1, one of two domains of primary structure in B-hordein polypeptides, which is also characterized by a series of degenerate tandem repeats (13). The short N-terminal domain that precedes the proline-rich repeats in other S-rich prolamins is absent from this B-hordein polypeptide (2). The boundary between domain 1 and domain 2, as defined from the earlier analysis of the cDNA sequences (13), is indicated in Fig. 2. Domain 1 of the pBHR184-encoded protein contains 79 residues which are 39% glutamine, 39% proline, 10% phenylalanine and include no sulphur amino acids. The repeated motif (based on the prototype octapeptide Pro-Gln-Gln-Pro-Phe-Pro-Gln-Gln) is evident throughout this domain, including the N-terminal 27 residues not previously sequenced (see ref. 2). The remaining 198 residues of the protein make up domain 2, which is distinguished from domain 1 by being relatively proline-poor, S-rich and non-repetitive (13). Domain 2 is 27% glutamine, 11% proline, 4.1% cysteine and 1.5% methionine.

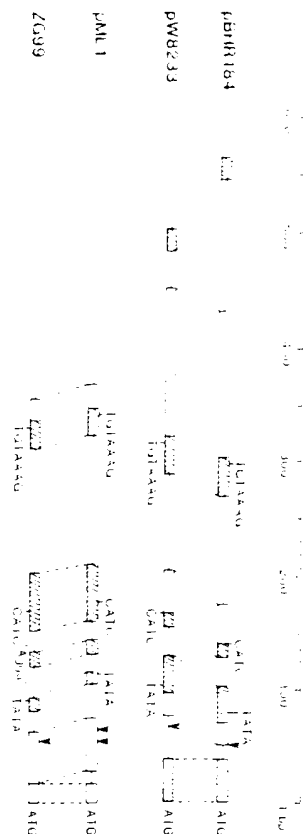


Fig. 4. Diagram showing the locations of conserved sequences in the 5'-flanking regions of two families of prolamin genes. The B1 hordein (pBHR104) and an α -glutadin gene from wheat (pW8233; ref. 30) are divergent representatives of the 5-rich prolamin multi-gene family, while the Z11 (pM1, ref. 45) and Z19 (Z699, ref. 25) genes are similarly divergent representatives of the zein multi-gene family. A graphic matrix homology plot was used to locate the most strongly conserved sequences in the 5'-flanking regions of each pair of genes and the sequences were then aligned manually to assess the extent of lower order homologies. The degree of homology is indicated by the shading within the rectangles. Diagonal shading: 80-92% identity; stippled: 74% identity; open: 56-62% identity. Sequences outside the rectangles show little or no homology (40%). Short 'core' sequences that characterize the most strongly conserved blocks of homology are indicated. The conserved 'Agda box' sequence that was noted in a survey of the promoter regions of a number of plant genes (48), including the zein gene family, was not found in the B1 hordein or α -glutadin genes. The 5'-flanking regions of three other α -glutadin genes have been sequenced (31,32) and do not differ by more than 9% from the pW8233 gene that was used for these comparisons. Upstream sequences of at least 210 bp have been determined for two other Z11 genes, and these are 85% (27; ref. 29) and 90% (Z41; ref. 27) homologous to pM1, and for three Z19 genes, which are 99% (Z4; ref. 26), 96% (ZE19; ref. 28) and 88% (ZE25; ref. 28) homologous to Z499.

Conserved sequences in the 5'-flanking region

On the assumption that sequences important in gene expression are likely to be conserved among a group of genes with the same pattern of expression (44), we have carried out a detailed comparison between the flanking sequences of the B1 hordein gene and those of an α -gliadin gene from wheat. Although they are related proteins, the B1 hordeins and the α -gliadins are among the most divergent forms of the S-rich prolamins (2). A diagrammatic representation of the sequence homologies upstream of the two S-rich prolamin genes is shown in Fig. 4, along with a similar comparison between two maize genes that code for polypeptides belonging to the tight chain (Z19) and heavy chain (Z21) classes of zein. Despite considerable divergence between the 5'-flanking sequences of the B1 hordein and α -gliadin

genes (<50% overall homology upstream of the cap sites) there are several short segments within 600 bp of the translation initiation codons that show more than 80% homology. Significantly, several of these most strongly conserved segments are related to sequences that are conserved at similar locations in the zein multi-gene family. Two of the sequences that are common to both multi-gene families are located at around -100 and -150 (relative to the ATG codon) and may be the counterparts of the TATA and CCAAT boxes that are components of the promoter region of animal genes (46, 47). A TATA box sequence has been found in almost all plant genes so far analysed (48), while a sequence similar to the CCAAT box-like segment in the prolamin genes is also found in a wide variety of other cereal genes (Fig. 5A) (although not in the maize *Adh2* gene (53)). The CCAAT box-like sequence has been designated a 'CATC' box by virtue of the most strongly conserved tetranucleotide within the 11 bp segment. Sequences conforming to the CATC box consensus (Fig. 5A) were not found in a manual survey of the corresponding region upstream of the TATA box in 15 published dicot gene sequences. Despite the absence of a clearly identifiable CCAAT box-like sequence that is common to both monocots and dicots (see also ref. 48), the importance of sequences in this region for maximal gene expression in plants has been demonstrated for two genes (54, 55).

Potentially the most interesting conserved sequences in the 5'-flanking regions of the two families of prolamin genes are found about 300 bp upstream of the ATG codon (Fig. 4). These sequences, or '-300 elements', are aligned in Fig. 5B to illustrate the features that are common to both multi-gene families. In the α -gliadin gene the -300 element is imperfectly repeated about 200 bp further upstream and in the B1 hordein gene at least part of the element is imperfectly repeated about 270 bp upstream (Fig. 5B). Sequences homologous to the -300 element were not found in the other cereal genes for which extensive upstream sequence data are available (49, 50, 52, 53).

An indication of the very low frequency of random occurrence of sequences related to the -300 element was obtained by searching the Genbank nucleotide sequence database (release 18.0, 3×10^6 nucleotides) for all occurrences of a 28 bp consensus sequence: ANNGTGAACGMAATATNGATG (where W = G or T, and invariant nucleotides are underlined). The consensus sequence was derived by aligning the -300 element and its repeats in the B1 hordein gene and in all published α -glutadin (30-32) and 221 zein (27, 29, 44) sequences (without introducing gaps). The 219

Clone	'CAAT box'	'TATA Box'	Cap Site(s)
pBRL184:	-139 ACAACCAACA.....	-80 CATATAATA.....	(-51).....
pW6233:	-165 GCATCAAGCA.....	-105 CATATAA.....	(-57).....
PML1:	-177 CATCATATAC.....	-112 CATATAA.....	(-65, -52).....
Z699:	-168 TCATCTTACC.....	-90 GATATAATA.....	(-57).....
PLS.1:	-167 CCATCTCTCC.....	-140 CATATAA.....	(-99).....
PTH012:	-201 CCATCTCAGC.....	-97 CATATTAC.....	(-62).....
pW64.3:	-113 CCATCCACAC.....	-79 CATATAA.....	(-42).....
Consensus:	CAATCAAC	CTATAA ^T _A A	
Animal Consensus:	GA ^T CAATCT	TATAA ^T _A	

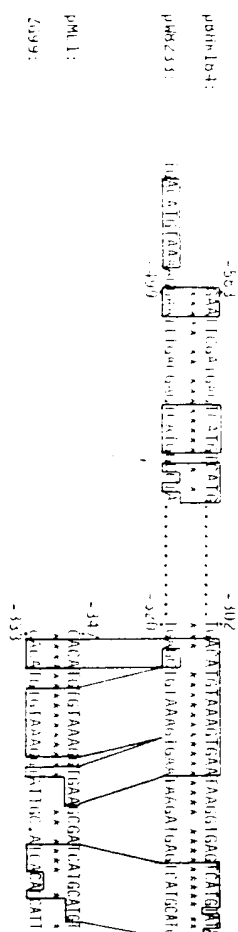


Fig. 5. Sequences common to the 5'-flanking regions of the B1 hordein (pBRL184), α -gliadin (pW6233), Z21 (PML1) and Z19 (Z699) genes. Sequences that are common to all four genes analysed in Fig. 4 have been aligned here for comparison. (A) Conserved sequences located within 200 bp upstream of the ATG codon. Similar sequences are found at corresponding positions in a variety of other cereal genes (49-52) and a representative selection of these is also shown. PLS.1: maize alcohol dehydrogenase 1 (Adh1) gene (50); PTH012: wheat H3 histone gene (51); pW64.3: wheat Rubisco small subunit gene (51). The consensus sequence for the TATA box is similar to that previously determined for a group of genes from both monocots and dicots (48). (B) Conserved distal sequences that are located at around position -300 in both families of prolamins and that are repeated further upstream in the α -gliadin gene and (at least partially) in the B1 hordein gene. The sequence of the repeat of the -300 element in the B1 hordein gene is incomplete because the repeat contains one of the two EcoRI sites that define the boundaries of the sequenced 2.9 kb fragment. Asterisks indicate identical residues in each pair of aligned sequences, while those parts of the element that are common to the S-rich prolamins and zein genes are boxed. Numbering is relative to the ATG codon. The corresponding sequences in other α -gliadin (31, 32), Z21 (27, 29) and Z19 (26, 28) genes are similar or identical to the representative sequences in this figure.

sequences were omitted because they deviate markedly from the consensus in the second half of the -300 element (see Fig. 5B). Even allowing up to three mismatches at the variant positions (the maximum deviation shown by

any of the sequences that were used to derive the consensus), no additional occurrences were found in the database. Therefore, in view of the apparently independent evolutionary origins of the S-rich prolamins and zein genes, it seems that the presence of this sequence at approximately the same location in both genes is highly unlikely to have arisen by chance. This suggests that the -300 element has some important function, which (in view of its location) is most likely to be related to the control of prolamins gene expression. Conserved 5'-flanking sequences have previously been noted in another group of developmentally co-regulated plant genes (56) and there is accumulating evidence that short upstream sequences are involved in the coordinate induction of unlinked eucaryotic genes (56-58). We are currently investigating the possibility that, by analogy with *Drosophila* heat shock genes (60), there is a specific interaction between DNA-binding proteins in endosperm nuclei and conserved upstream sequences in the B1 hordein gene.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. B. J. Miflin for many helpful discussions throughout the course of this work and for critical reading of the manuscript. The research was supported partly by contract no. 601-4-023-UK(H) of the Biomolecular Engineering Programme of the Commission of the European Communities.

REFERENCES

- Shewry, P.R., Miflin, B.J. and Kasarda, D.D. (1984) Phil. Trans. R. Soc. Lond. B 304, 297-308.
- Kreis, M., Shewry, P.R., Forde, B.G., Forde, J. and Miflin, B.J. (1985) Oxf. Surveys Plant Molec. Cell Biol. 2, 253-317.
- Greene, F.C. (1983) Plant Physiol. 71, 40-46.
- Rahman, S., Shewry, P.R. and Miflin, B.J. (1982) J. Exp. Bot. 33, 717-728.
- Mitra, P.S. and Mertz, E.T. (1975) Cereal Chem. 52, 734-739.
- Rahman, S., Shewry, P.R., Forde, B.G., Kreis, M. and Miflin, B.J. (1983) Planta 159, 366-372.
- Thompson, R.D. and Bartels, D. (1983) Plant Sci. Lett. 29, 295-304.
- Kreis, M., Shewry, P.R., Forde, B.G., Rahman, S., Bahramian, M.B. and Miflin, B.J. (1984) Biochem. Genet. 22, 231-255.
- Manzocchi, L.A., Dammati, M.G. and Gentilella, E. (1980) Maydica 25, 199-210.
- Soave, C. and Salami, F. (1983) in Seed Proteins, Daussant, J., Mosse, J. and Vaughan, J. Eds., pp. 205-218, Academic Press, London.
- Benton, W.D. and Davis, R.W. (1977) Science 196, 180-182.
- Rigby, P.W.J., Dieckmann, M., Rhodes, C. and Berg, P. (1977) J. Mol. Biol. 113, 237-251.

- 13 Forde, B.S., Kreis, M., Williamson, M.S., Fry, R.P., Pywell, J., Shewry, P.R., Bunce, N. and Mifflin, B.J. (1985) EMBO J. 4, 9-15.
- 14 Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory, New York.
- 15 Clewell, D.B. and Hellmuth, D.R. (1970) *Biochemistry* 9, 4428-4440.
- 16 Guo, L.H., Yang, R.C.A. and Wu, R. (1983) *Nucl. Acids Res.* 16, 5521-5540.
- 17 Messing, J. (1983) *Meth. Enzymol.* 101, 20-78.
- 18 Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463-5467.
- 19 Staden, R. (1982) *Nucl. Acids Res.* 10, 2951-2961.
- 20 Staden, R. (1982) *Nucl. Acids Res.* 10, 4731-4751.
- 21 Staden, R. (1984) *Nucl. Acids Res.* 12, 505-519.
- 22 Berk, A.J. and Sharp, P.A. (1977) *Cell* 12, 721-732.
- 23 Southern, E.M. (1975) *J. Mol. Biol.* 98, 503-517.
- 24 Rasmussen, S.K., Hopp, H.E. and Brandt, A. (1983) *Carlsberg Res. Commun.* 48, 187-199.
- 25 Pedersen, K., Beveraux, J., Wilson, D.R., Sheldon, E. and Larkins, B.A. (1982) *Cell* 29, 1015-1026.
- 26 Hu, N.T., Pelter, M.A., Heidecker, G., Messing, J. and Rubenstein, I. (1982) *EMBO J.* 1, 1337-1342.
- 27 Speria, A., Viotti, A. and Pirotta, V. (1983) *EMBO J.* 2, 1589-1594.
- 28 Speria, A., Viotti, A. and Pirotta, V. (1983) *J. Mol. Biol.* 169, 799-811.
- 29 Kril, J., Vieira, J., Rubenstein, I. and Messing, J. (1984) *Gene* 28, 113-118.
- 30 Rafalski, J.A., Scheets, K., Metzler, M., Peterson, D.M., Hedgcock, C. and Solt, D.G. (1984) *EMBO J.* 3, 1409-1415.
- 31 Anderson, O.D., Litts, J.C., Gautier, M.-F. and Greene, F.C. (1984) *Nucl. Acids Res.* 12, 8129-8144.
- 32 Summer-Smith, M., Rafalski, J.A., Sugiyama, T., Scoll, M. and Solt, D. (1985) *Nucl. Acids Res.* 13, 3905-3916.
- 33 Proudfoot, N.J. and Brownlee, G.G. (1976) *Nature* 263, 211-214.
- 34 Matthews, J.A. and Mifflin, B.J. (1980) *Planta* 149, 262-268.
- 35 Forde, B.G., Kreis, M., Bahramian, M.B., Matthews, J.A., Mifflin, B.J., Thompson, R.D., Bartels, D. and Flavell, R.B. (1981) *Nucl. Acids Res.* 9, 6689-6707.
- 36 Watson, M.E.E. (1984) *Nucl. Acids Res.* 12, 5145-5164.
- 37 Cameron-Millis, V., Ingversen, J. and Brandt, A. (1978) *Carlsberg Res. Commun.* 43, 91-102.
- 38 Cameron-Millis, V. (1980) *Carlsberg Res. Commun.* 45, 557-576.
- 39 Brandt, A. and Ingversen, J. (1978) *Carlsberg Res. Commun.* 43, 451-469.
- 40 Shewry, P.R., Field, J.M., Kirkman, M.A., Faulks, A.J. and Mifflin, B.J. (1980) *J. Exp. Bot.* 31, 393-407.
- 41 Schmitt, J.M. and Svendsen, I. (1980) *Carlsberg Res. Commun.* 45, 143-148.
- 42 Kasarda, D., Okita, T.W., Bernardin, J., Baecker, P.A., Nimmo, C., Lew, E., Dietler, M. and Greene, F.C. (1984) *Proc. Natl. Acad. Sci. USA* 81, 4712-4716.
- 43 Von Heijne, G. (1983) *Eur. J. Biochem.* 133, 17-21.
- 44 Davidson, E.H., Jacobs, H.T. and Britten, R.J. (1983) *Nature* 301, 468-470.
- 45 Langridge, P. and Felix, G. (1983) *Cell* 34, 1015-1022.
- 46 Benoit, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980) *Nucl. Acids Res.* 8, 127-142.
- 47 Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., De Kriel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulters, C.C. and Proudfoot, N.J. (1980) *Cell* 21, 653-668.
- 48 Messing, J., Geraghty, D., Heidecker, G., Hu, N.T., Kril, J. and Rubenstein, I. (1983) in *Genetic engineering of plants*, Kosuge, T., Meredith, C.P. and Hollaender, A. Eds., pp. 211-227, Plenum Press, New York.
- 49 Dennis, E.S., Gerlach, M.L., Pryor, A.J., Bennett, J.L., Inglis, A., Llewellyn, D., Sachs, M.M., Ferl, R.J. and Peacock, W.J. (1984) *Nucl. Acids Res.* 12, 3983-4000.
- 50 Tabata, T., Fukasawa, M. and Iwabuchi, M. (1984) *Mol. Gen. Genet.* 196, 397-400.
- 51 Broglie, R., Coruzzi, G., Lampia, G., Keith, B. and Chua, N.H. (1983) *Biotechnology* 1, 55-61.
- 52 Tabata, T., Sasaki, K. and Iwabuchi, M. (1983) *Nucl. Acids Res.* 11, 5865-5875.
- 53 Dennis, E.S., Sachs, M.M., Gerlach, M.L., Finnegan, E.J. and Peacock, W.J. (1985) *Nucl. Acids Res.* 13, 727-743.
- 54 Shaw, C.H., Carter, G.H., Watson, M.D. and Shaw, C.H. (1984) *Nucl. Acids Res.* 12, 7831-7846.
- 55 Odell, J.T., Nagy, F. and Chua, N.H. (1985) *Nature* 313, 810-812.
- 56 Mauro, V.P., Nguyen, T., Katinakts, P. and Verma, D.P.S. (1985) *Nucl. Acids Res.* 13, 239-249.
- 57 Pelham, H.R.B. (1982) *Cell* 30, 517-528.
- 58 Donahue, T.F., Daves, R.S., Lucchini, G. and Fink, G.R. (1983) *Cell* 32, 89-98.
- 59 Miller, A.M., Mackay, V.L. and Nasmyth, K.A. (1985) *Nature* 314, 598-603.
- 60 Wu, C. (1984) *Nature* 311, 81-84

Conservation and variability of wheat α/β -gliadin genesMartin Sumner-Smith¹, J. Anton Radakovic², Lakshmi Srinivasan, Marie Stoll³ and Dieter Sell¹Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA

Received 28 March 1985; Accepted 6 May 1985

ABSTRACT

We have sequenced two genomic clones for wheat α/β -gliadin storage protein genes. Comparison with a known sequence reveals close homology between the three and confirms the previously suspected evolutionary relatedness of members of this gliadin family. The coding region can be divided into six domains. Two unusual structures were found within this region: (i) The P-boxes which are composed of 12 codons, six of which are for proline, that are tandemly repeated four or five times; and (ii) Two polyglutamine stretches which consist of 18-22 tandemly repeated glutamine codons in one case, and 7-28 in the second. Analysis of the P-box structures revealed that certain mutations were probably present in the hypothetical ancestral α/β -gliadin gene prior to gene multiplication. None of the genes have introns. All of the genes appear to contain typical eukaryotic promoters and also possess the double polyadenylation signal of plants.

INTRODUCTION

During wheat seed development the predominant protein synthesis is of two groups of proteins, totalling more than 50 members, which are thought to provide a stored source of nitrogen for future germination. These storage proteins, the gliadins and glutenins, have been the subject of extensive study.

Originally, the gliadins were classified according to their electrophoretic mobility in starch gels in aluminum lactate (1). Recently, they have been reclassified according to the size, amino acid composition and N-terminal sequences of purified species, into the predominant sulfur-rich α/β - and γ -gliadins, and the less abundant sulfur-poor omega-gliadins (2,3). Gliadins are very rich in glutamine (approx. 35%) and proline (approx. 15%).

On the basis of the apparent homology between the N-terminal sequences of purified members of each gliadin class, it is thought that the gliadins are the products of several multigene families (3-5). These families presumably arose by the repeated duplication of a few ancestral genes. Since all of the multigene families seem to be present in each of the ancestral genomes which have contributed to modern hexaploid wheat (6-8), multiplication of the original

gliadin genes must have occurred in some ancestor common to the diploid strains. Three gliadin gene loci have been identified by genetic means, two maps to the short arm of homoeologous chromosome group 1 and one to group 6 (9-14). Individual genes within these loci are tightly linked (11-17).

Wheat storage proteins represent a convenient system to study both the coordinate expression of several gene families during development and also the evolution of these families. We have previously presented the complete sequence of an α/β -gliadin gene and its flanks (18). Here we report the sequence of two additional genomic clones, discuss a domain structure for gliadin proteins and attempt insights into the evolution of their genes.

MATERIALS AND METHODS

Materials: Klenow fragment of *E. coli* DNA polymerase I was a gift of Harry Templeton, Yale University. Other materials were obtained commercially.

General: Handling and analysis of nucleic acids, including restriction enzyme digestions, agarose electrophoresis and elution, Southern blots, ligations, plasmid and phage DNA isolation were by established methods (19). Bacterial transformation was by the method of Hanahan (20).

Gliadin Clones: The gliadin genes selected here were selected from a wheat (cv. Yaman) partial EcoRI library in Charon 32 (courtesy of Drs. J. Slight and M. Murray, Agrigenetics Corp., Madison, WI) and recloned into pBR325 (21) in both orientations. The complete sequence of pW8233 has already been described (18). Restriction maps were determined for pW8142 and pW1215 (now shown) and the gliadin gene localized by Southern hybridization. pW8142 contains a 7.7 kb fragment, within which is a 3231 bp EcoRI-HindIII subfragment carrying the gene. pW1215 has a 9.8 kb fragment within which a 3043 bp HindIII-HindIII subfragment contains the gene. The two subfragments were completely sequenced. In addition, the ends of the flanking subfragments were sequenced. Data not shown in Fig. 1 were submitted to GenBankIm.

Sequencing: Both strands of the gliadin gene containing subfragments of pW8142 and pW1215 were determined by the dideoxy method using M13mp8, M13mp9, M13mp10 and M13mp11 and DNA fragments generated by a variation of the Bal31-deletion method (22,23). Sequence data were compiled and analyzed using the programs described by Larson and Messing (24) and by Sege et al. (25).

Nuclease S1 Mapping of the 5'-End of α/β -Gliadin mRNA: For all three genes examined, complementary probes (coding strand) were made by universal oligonucleotide (17-mer) primed synthesis on appropriate Bal31-deleted templates cloned in M13. Sequencing conditions were used except that dideoxynucleotides

were omitted. The resulting partially double-stranded molecules were cleaved with PstII (after nucleotide 653, see Fig. 1). Restricted probes were recovered after phenol extraction, denatured by boiling in 50% formamide and hybridized overnight with an excess (10 ng) of wheat endosperm poly(A) RNA (courtesy of K. Scherfs, Kansas State University) according to the conditions (80% formamide, 0.4 M NaCl, 0.04 M Pipes, pH 6.4, 1 mM EDTA) of Weaver and Weissmann (29) at 53° under paraffin oil. Controls contained no poly(A) RNA. After hybridization, samples were added to 200 μ l 51 buffer (0.25 M NaCl, 30 mM sodium acetate pH 4.6, 1 mM ZnSO₄, 20 ng/ μ l denatured salmon sperm DNA) and incubated at 30° M. DNA was recovered by ethanol precipitation and run on an 8% sequencing gel alongside a set of sequencing reactions as a length reference.

RESULTS AND DISCUSSION

Gene Sequences: The sequences of the three genes and their immediate flanks are shown in Fig. 1. The predicted N-terminal amino acid sequences and compositions are consistent with those of α/β -gliadin genes (3). The genes are clearly related, but show mutational differences at a number of sites as well as changes which could have arisen by insertions and deletions that preserve the reading frames.

The 5'-flanks of the genes are homologous for approximately 600 bp upstream of the ATG start codon. The sequences diverge 20-30 bp upstream of the HindIII site (Position 1 in Fig. 1). Clones pW8142 and pW1215 share homologous 3'-flank for at least 1600 bases (data not shown), but these are not related to the 3'-flank of pW8233 beginning at the position 1680. Comparison with 3'-noncoding regions of barley B1 hordein cDNA clones (30) revealed a close homology that extends downstream from the translation stop codon to the second polyadenylation signal. The spacing of this polyadenylation signal and of the stop codon is also conserved between the hordein clones and pW8233/pW8142 (except for a two bp deletion). This homology complements the observed similarity in the coding sequences of hordein and gliadin genes (see below and refs. 30-32). Close homology of zein genes has been reported (26, 27).

The ends of the mRNA: All three genes possess a typical eukaryotic RNA polymerase II promoter sequence (TATATAA/TA) 104 bases upstream of the ATG start codon. We determined the 5'-end of the α/β -gliadin mRNA species by nuclease S1 protection studies (Fig. 2). Subtracting 1-2 bases to account for the putative 5'-cap, we estimate that transcription *in vivo* begins 30 bases downstream from the end of the TATA-box at the indicated A (Fig. 1). The 3'-flank contains two potential polyadenylation signals (AATAAA/T and AATAAA)

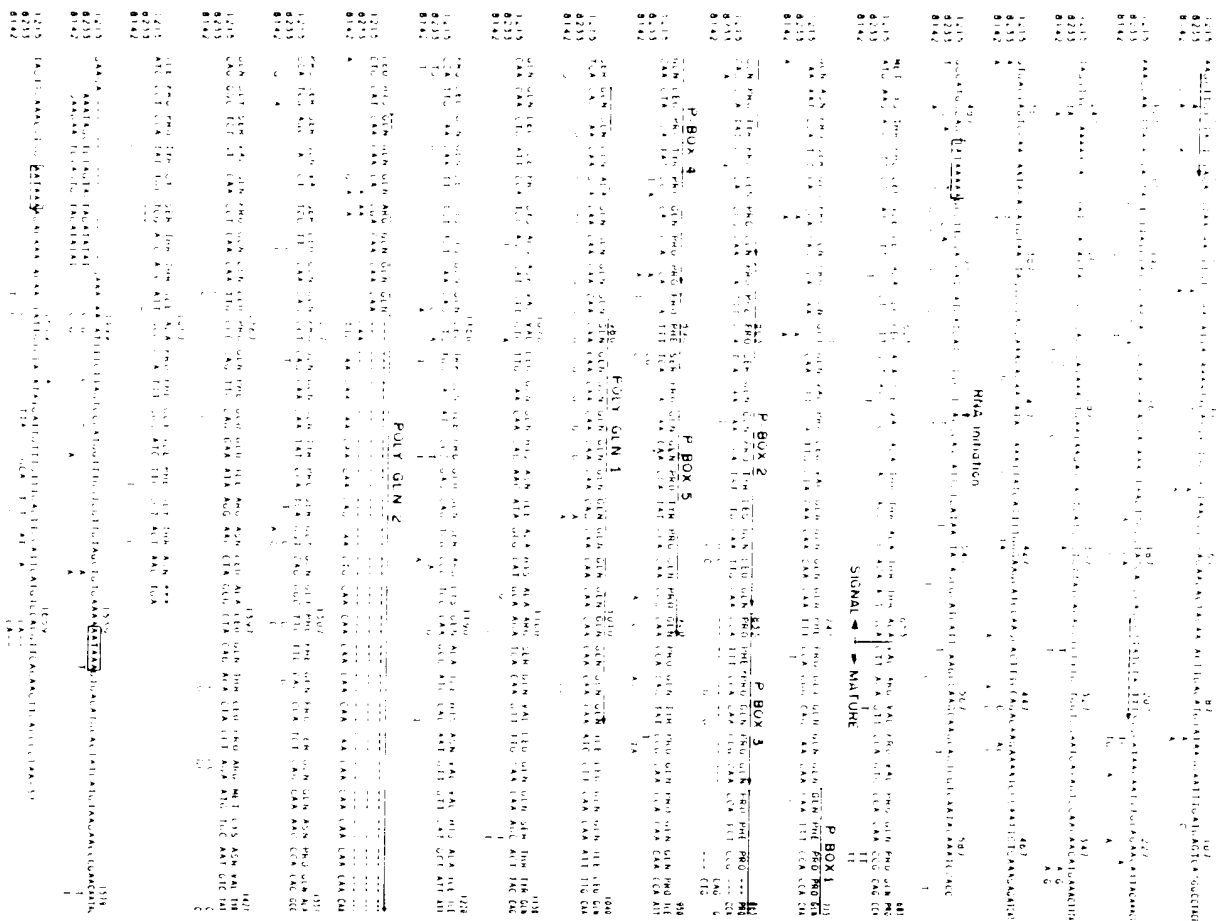


Figure 1. The DNA sequence and the derived protein sequence of three α/β -glutadin genes. The sequenced region of the clone pM1215 is presented (upper line); bases in clones pM823 and pM8142 which differ from those are indicated below; deleted bases are indicated by a dash. The numbering is

based on the pM1215 sequence. The predicted amino acid sequence of the product of pM1215 is indicated. The position of the P-boxes (see text) is indicated by arrows. The polyglutamine stretches. The direct repeat in the 5'-flank (see text) is indicated by arrows. The TATA box and polyadenylation signals are indicated. The RNA initiation (arrow) and the polyadenylation (*) site are given.

(Fig. 1); sequencing of four cDNA clones shows that the poly A tail starts at position 1625 (Fig. 1) (18).

Domain Structure of the Coding Region: Fig. 3 shows a generalized structure of wheat α/β -gliadin genes derived from the sequences of the three genomic clones. The coding region can be divided into six domains: a signal peptide, a region of nine dodecapeptide repeats, five of which show very close homology (the P-boxes), two polyglutamine stretches and two regions of unique sequence. A similar structure has recently been proposed (31); however, since we have a larger number of sequences to compare we observe more detail in this structure. The repeat regions and the polyglutamine stretches are further discussed below.

P-box: We have derived a consensus sequence for the 12 codon repeat (Fig. 4). This sequence, which might represent the ancestral gene, is the sequence from which the fewest base changes (mutations) are necessary to arrive at the actually observed sequences. Six of the 12 codons in the consensus sequence are for proline and represent the greatest density of these codons in the genes; therefore the designation P-box. There are also four glutamine codons and one codon each for tyrosine and phenylalanine. While all but one each of the proline and glutamine codons show mutational changes in one or more P-box examples, both the tyrosine and phenylalanine codons are unchanged in any box. The phenylalanine and tyrosine codons are exactly half the box size apart (i.e. six codons). This periodicity is preserved by three deletion variants of the third P-box, i.e. by deletion of exactly half of the box in two cases (pW8142, pW1215) and the entire box in one (pW8233). The periodicity is disrupted slightly in one case by the insertion of one codon in the fourth P-box of pW8233.

The P-box presumably arose in the ancestral α /3-gliadin gene and was multiplied prior to the extensive multiplication of the whole gene. Base changes are present in every sample of the P-box in the genes described here; no single box corresponds exactly to the consensus sequence. In some cases these mutations are present in all of the examples of the P-box at a given position, e.g. the A to T mutation in the sixth codon of Box 4. These 'early' mutations presumably arose in a particular P-box in the ancestral gene (or at least in an ancestor to all of the genes described here), and were preserved during

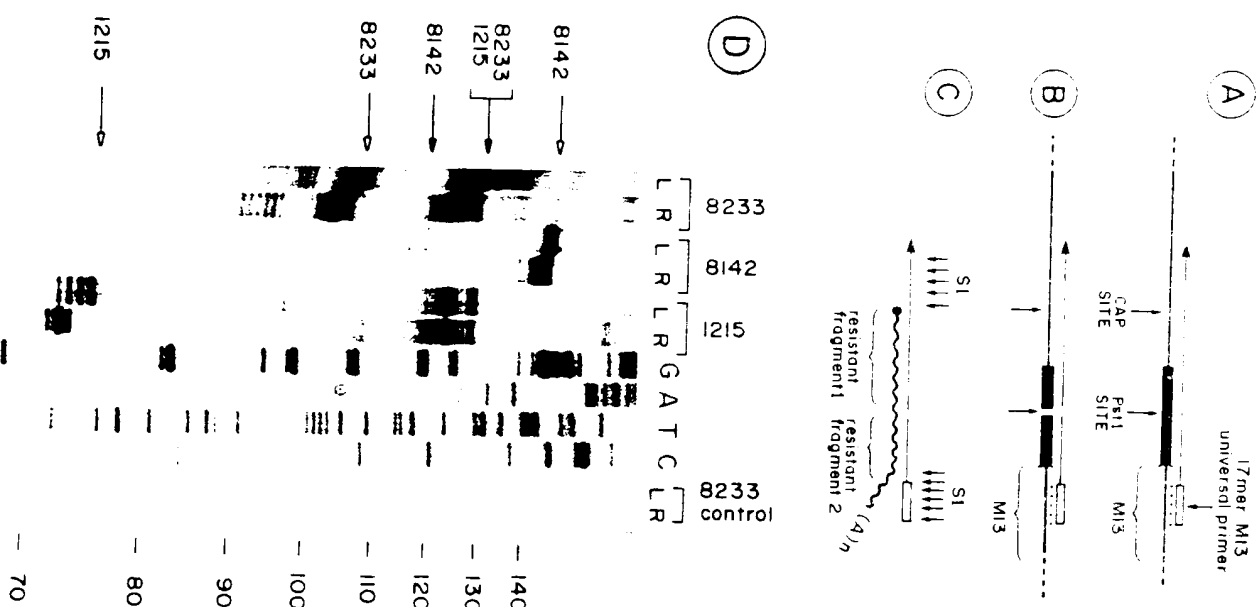


Figure 2. Nuclease S1 mapping of the 5'-end of α/β -gliadin mRNA. DNA complementary to mRNA was synthesized from appropriate BAL-31 deletion clones in M13 from each of the genes (A, top legend) and cleaved by PstI to produce uniform 3'-end (B). Two fragments of this DNA were protected from digestion by nuclease S1 (reaction times 5 min [lanes L] and 40 min [lanes R]) after hybridization to mRNA (C,D): 1.) The first fragment extends from the 5'-end of mRNA to the Pst site (D, filled arrows). The length of this fragment establishes the location of the cap site, in basepairs, upstream from the PstI site (position 653). 2.) The fragment extends from the PstI site to the end of the deleted clone (D, open arrows). The length of this fragment was different for each clone, because the deletions used were different. The sizes of the fragments were determined by comparison to sequencing reactions performed on a known clone run in adjacent lanes; calculated lengths are given in the right hand column of D. The size of the 5'-fragment from pm8142 was nine bases smaller than those for the other two genes because of a three codon deletion in the signal peptide of that gene. Control reactions, lacking mRNA were performed for all three clones - that for the pm8233 subclone is shown. The origin of the faint bands in L is unclear. For further details see Experimental Section.

gene multiplication. It might be argued that these mutations could have arisen in one gene of the repeated family and been spread to the others by unequal crossing over; however such a process would be more likely to transfer a given mutation to other P-boxes within the same gene, and there is little evidence for this.

Recently, the sequences of two complete α/β -gliadin cDNA clones (pG1A-42 and pG1A41) from different cultivars of *T. aestivum* were elucidated (31,33). Comparison of these sequences in the P-box region reveals that the seven early mutations in the five boxes are present as expected (Fig. 4). Further, amino acid sequence analysis (31) of a mixed α -gliadin fraction confirms that the four non-silent 'early' mutations are present in the five or more polypeptides which make up that fraction.

The significance of the P-box organization is unknown at this time. Possibly these peptides may confer on the proteins a structure important for their function *in vivo*. A γ -gliadin cDNA sequence containing the 5'-half of the mature coding region shows 14 repeats of a 7 codon sequence (28). In the case of corn, certain zein genes show a 70 codon sequence tandemly repeated nine times that make up the bulk of the final polypeptide (26,27,34). It has been suggested that each example of this repeat is able to assume an α -helical structure and that the nine resulting helices are able to stack side-by-side (35).

The Polyglutamine Stretches: A second interesting feature is the presence of two long polyglutamine stretches (Figs. 1 and 4). The first one is found near the center of each gene and consists of from 18-22 codons. In two of the genes this stretch is composed of a single CAG codon followed by either 20

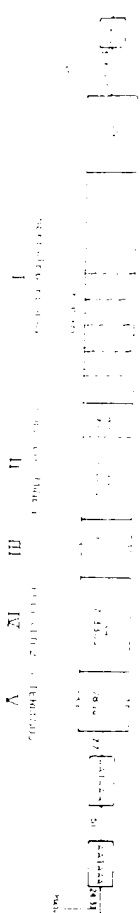


Figure 3. General structure of α/β -gliadin genes. The generalized structure is derived from the three genomic clones described here as well as from the cDNA clones described by Kasarda et al. (31) and Proffitt et al. (30). The TATA box, CAP site, polyadenylation signals and site are found in the indicated number of bases away from the coding sequence shown as an enlarged box. The signal sequence precedes five regions in the mature polypeptide defined by analysis of the cDNA sequences. The first region (I) consists of a series of 9, typically double-stranded repeats. Five of these repeats (crosshatched) appeared to be more closely related to each other (P-boxes, Fig. 4). The first repeat, which is putative, is preceded by a three codon stretch having no obvious relationship to the rest of the region. Two polyglutamine stretches (II and IV) separate two regions of non-repeated sequence (V).

(pW8142) or 21 (pW1215) CAA codons, whereas in the other the stretch consists of nine CAG codons followed by nine CAA codons (pW8233). These stretches are actually disrupted by a putative mutant GCA codon at the fourth position in both pW8142 and pW1215, as well as a silent A to G mutation in the fifteenth position of pW1215. It therefore appears that the polyglutamine stretch was initially generated by multiplication of a CAA codon and subsequently in some genes (e.g. pW8233) by multiplication of a CAG codon found immediately in front of the CAA stretch. The latter event might therefore be more recent. Alternatively, the poly CAG stretch may have been reduced to a single codon in some genes. Another possible mechanism to account for the CAG to CAA transition derives from the fact that CG and CHG sequences in wheat DNA are over 80% methylated to m5C (36). Should 5-methylcytosine suffer spontaneous deamination to an appreciable extent (37), CAG would tend to be converted to CAA. Deamination of C in the first position of CAG would result in a nonsense codon. Thus, a selective change of CAG to CAA could be explained.

pW8142 shows a second polyglutamine stretch later in the gene consisting of 28 codons, 3 of which are mutated away from glutamine. A similar but shorter stretch is located at the corresponding position (base 1227 in pW1215) in the other two genes described here, which consists of 7 (pW1215) or 8 (pW8233) codons, one of which is mutated away from glutamine. In pG11A-42 this stretch has 33 glutamine codons (31).

Evolutionary Implications: Since the α/β -gliadin genes in the three diploid genomes which contribute to the hexaploid genome of modern wheat are present

Consensus Sequence	Pro	Phe	Pro	Pro	Gln	Gln	Pro	Tyr	Pro	Gln	Pro	Gln
	CCA	TTC	GCA	GAA	CAA	CAA	TAA	ATG	ATG	CAA	ATG	CAA
Box 1	-A-	---	---	---	-g-	---	---	---	---	---	---	---
	-A-	---	---	---	-g-	---	-a	---	---	---	---	---
	-A-	---	---	---	-q	---	---	---	---	---	---	---
	---	---	---	---	-q	---	-a	---	---	---	---	---
Box 2	---	---	T-	---	---	---	-T-	-a	TT-	---	---	---
	---	---	T-	---	-T-	---	-T-	---	-T-	---	---	---
	---	---	T-	---	---	---	-T-	---	-T-	---	---	---
	-t	---	T-	---	---	---	-T-	---	-T-	---	---	---
Box 3	---	---	T-	---	---	---	-Ta	---	-T-	---	---	---
	---	---	---	---	---	---	---	---	---	---	---	---
	---	---	---	---	---	---	---	---	---	---	---	---
	---	---	---	---	---	---	---	---	---	---	---	---
Box 4	---	---	-g	---	-T-	---	---	---	-a	-C-	---	---
	---	---	-g*	-g	-T-	---	T-a	---	-a	---	---	---
	---	---	-Tg	---	-T-	---	---	---	---	---	---	---
	---	---	-g	---	-T-	---	---	---	---	---	---	---
Box 5	---	---	-g*	-g	-T-	---	T-	---	-a	---	---	---
	---	---	T-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	T-	---	---	---	-a	-a	-G-	---	---	---
	---	---	-G-	---	---	---	-a	-a	-A-	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---	-a	-a	---	---	---	---
	---	---	-G-	---	---	---						

Figure 4. Putative mutations in the P-boxes. The P-boxes within each gene are compared to the consensus sequence derived from all of the boxes (upper line). The sequential position of the five boxes indicated in Fig. 1 is shown. Putative mutations, i.e. bases which differ from the consensus sequence are indicated; silent mutations are shown in lower-case characters, whereas mutations which change the coded amino acid are shown in upper-case letters. Bases which are preserved are indicated by a horizontal line. Deletions are indicated by a blank. * indicates insertion of a CAG codon. Data from *T. aestivum* cDNA clones pG11A-42 (31) and pCH1941 (33) are also included.

in multiple copies (6-14), the multiplication of that gene is likely to have occurred before these species diverged. This hypothesis could be confirmed if the mutations which we believe occurred in the ancestral gene are found in the α/β -gliadins derived from each diploid genome. At present we do not know from which genome the genes which we have sequenced have been derived. Analysis of DNA from the diploid ancestors of wheat is needed.

The high number of mutations in the P-boxes implies that they are evolutionarily ancient structures; however, we cannot yet derive an estimate of the rate of accumulation of mutations in the structure. The gliadin genes of wheat share homology with various hordein genes from barley (40,32). Thus, we would expect that the P-box structure and some of the earliest mutations in it might be preserved in barley hordeins. If so, it may then be possible to calibrate the rate of mutant accumulation using the time of divergence of wheat ancestors and barley as a standard, and thus calculate the evolutionary age of various features of the genes.

It is possible to make a crude estimate of the relative period of time that the repeated P-box structure existed prior to the multiplication of the ancestral gene, compared to the time since that event first occurred. Since the number of P-box mutations common to all genes is approximately equal to the number of mutations unique to any one gene, one can speculate that the repeated P-box structure must have existed for about as long before gene multiplication began as after.

Possible Regulatory Sequences: The expression of gliadin genes is coordinately regulated during seed development (38). It is therefore likely that these genes share common target sequences for the developmental regulatory mechanism(s). Such sequences have been demonstrated in some animal and viral genomes, and typically are present in multiple copies in the flanking regions of the regulated genes (39). If this generalization is true for plants we would expect the sites to be in the large stretches of conserved sequence in the 5'- and/or 3'-flanking regions of these genes. Several repeated sequences were found in the common 5'-flanks. One of these, which is present twice in the 5'-flank (Fig. 1), shows a surprising homology (up to 73%) to a 19 bp consensus sequence flanking the ovalbumin and related genes and identified as a probable binding site for chicken oviduct progesterone receptor (40). There are no common shorter sequences in the non-homologous 3'-flanks of PM8142/PM1215 and PM8233 (data not shown).

ACKNOWLEDGEMENTS

We are indebted to Jerry Slighum and Michael Murray for the gift of a wheat genomic library and to Nancy Templeton for *E. coli* DNA polymerase. We are grateful to Thomas Okita and Ralph Quatrano for making their sequence data available to us before publication.

M.S.-S. was a Fellow of the Medical Research Council of Canada; A.R. was

supported by E.I. du Pont de Nemours & Co.; T.S. was supported by a grant from the National Institutes of Health.

Present address: Alkermes, Inc., 6850 Gateway Drive, Massachusetts, Ontario L4V 1P1, Canada
Present address: Central Research and Development Department, Experimental Station, E.I. du Pont de Nemours & Co., Wilmington, DE 19898, USA
Permanent address: Division Genetech, Institut de Recherches Biologiques C. Esdaile, Montreal, Quebec

REFERENCE

1. Mowchik, J.H., Boundy, J.A., and Philter, R.J. (1961) *Arch. Biochem. Biophys.* **94**, 477-482.
2. Kasarda, D.D., Autran, J.-C., Lew, E.J.-L., Nimmo, C.C., and Shewry, P.R. (1983) *Biochim. Biophys. Acta* **747**, 138-150.
3. Bietz, J.A., Hedener, F.R., Sanderson, J.E., and Wall, J.S. (1977) *Cereal Chem.* **54**, 1070-1083.
4. Kasarda, D.D., DaRosa, D.A., and Ohms, J.I. (1974) *Biochim. Biophys. Acta* **351**, 290-294.
5. Shewry, P.R., Autran, J.-C., Nimmo, C.C., Lew, E.J.-L., and Kasarda, D.D. (1980) *Nature* **286**, 520-522.
6. Lawrence, G.J., and Shepherd, K.W. (1981) *Theor. Appl. Genet.* **60**, 333-337.
7. Wrigley, C.W., Lawrence, G.J., and Shepherd, K.W. (1982) *Aust. J. Plant Physiol.* **9**, 15-30.
8. Gallit, G., and Feldman, M. (1983) *Theor. Appl. Genet.* **66**, 77-86.
9. Wrigley, C.W., and Shepherd, K.W. (1973) *Ann. NY Acad. Sci.* **209**, 154-162.
10. Kasarda, D.D., Bernardin, J.E., and Qualset, C.O. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 3646-3650.
11. Brown, J.W.S., Kemble, R.J., Law, C.N., and Flavell, R.R. (1979) *Genetics* **93**, 189-200.
12. Brown, J.W.S., Law, C.N., Worland, A.J., and Flavell, R.R. (1981) *Theor. Appl. Genet.* **59**, 361-371.
13. Gallit, G., and Feldman, M. (1983) *Theor. Appl. Genet.* **64**, 97-101.
14. Gallit, G., and Feldman, M. (1984) *Mol. Gen. Genet.* **193**, 293-298.
15. Damidaux, R., Autran, J.-C., Gagnier, P., and Feillet, P. (1980) *C.R. Acad. Sci. (Paris)* **291D**, 585-588.
16. Sozinov, A.A., and Poyereida, F.A. (1990) *Ann. Technol. Agric.* **29**, 229-245.
17. Payne, P.I., Jackson, E.A., Holt, T.M., and Law, C.N. (1984) *Theor. Appl. Genet.* **67**, 235-243.
18. Rafalski, J.A., Scheeels, K., Metzler, M., Peterson, D.M., Hedgcock, C., and Solt, D.G. (1984) *EMBO J.* **3**, 1409-1415.
19. Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) *Molecular Cloning - A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
20. Hanahan, D. (1983) *J. Mol. Biol.* **166**, 557-580.
21. Bolivar, F. (1978) *Gene* **4**, 121-136.
22. Guo, L.-H., Yang, R.C.A., and Wu, R. (1983) *Nucl. Acids Res.* **11**, 5521-5540.
23. Poncz, M., Solowiejczyk, D., Ballantine, M., Schwartz, E., and Surrey, S. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 4298-4302.
24. Larson, R., and Messing, J. (1982) *Nucl. Acids Res.* **10**, 39-49.
25. Sege, R.D., Solt, D., Ruddle, F.H., and Queen, C. (1981) *Nucl. Acids Res.* **9**, 437-444.
26. Geraghty, D., Peifer, M.A., Rubenstein, I., and Messing, J. (1981) *Nucl. Acids Res.* **9**, 5163-5174.

27. Messing, J., Bergshly, D., Heidecker, G., Hu, H.-I., Kridl, J., and Rubenstein, L. (1983) In: *Genetic Engineering of Plants: An Agricultural Perspective* (Kosuge, T., Meredith, C.P., Hollander, A., eds.) pp. 211-227 (Plenum Press, NY).
28. Schect, K., Rafalski, J.A., Hedgock, C., and Soli, D.G. (1985) *Plant Sci. Letters*, in press.
29. Weaver, R.F. and Weissmann, C.H. (1979) *Nucl. Acids Res.* **7**, 1175-1193.
30. Kasmussen, S.K., Hupp, E., and Brandt, A. (1983) *Carlsberg Res. Commun.* **48**, 187-199.
31. Kasarda, D.D., Okita, T.W., Bernardin, J.E., Baecker, P.A., Nimmo, C.C., Lew, E.J.-I., Dieler, M.D., and Greene, F.C. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 4712-4716.
32. Forde, B.G., Kreis, M., Bahramian, M.B., Matthews, J.A., Mifflin, B.J., Thompson, R.D., Bartels, D., and Flavell, R.B. (1981) *Nucl. Acids Res.* **9**, 6639-6707.
33. Proffitt, J.H., Chakerian, K.L., Sheehy, R.E., and Quatrano, R.S., manuscript submitted.
34. Pedersen, K., Devereux, J., Wilson, D.R., Sheldon, E., and Larkins, B.A. (1982) *Cell* **29**, 1015-1026.
35. Argos, P., Pedersen, K., Marks, M.U., and Larkins, B.A. (1982) *J. Biol. Chem.* **257**, 9984-9990.
36. Gruenbaum, L., Naveh-Baron, Cedar, T., and Razin, A. (1981) *Nature* **292**, 860-862.
37. Lindahl, T. and Hyberg, B. (1974) *Biochemistry* **13**, 3405-3410.
38. Kasarda, D.D., Bernardin, J.E., and Nimmo, C.C. (1976) In: *Advances in Cereal Science and Technology*, L. Pomeroy, Y., ed., Am. Assoc. of Cereal Chemists, St. Paul, MN.
39. Davidson, E.H., Jacobs, H.T., and Britten, R.J. (1983) *Nature* **301**, 468-470.
40. Mulvihill, E.K., Lefevre, J.-P., and Chambon, P. (1982) *Cell* **28**, 621-632.

(Cloning of cDNA sequences for an *Artemia salina* hnRNP protein: evidence for conservation through evolution)

Madhu Chuz Alvarez, Madhomer Szol and Angel Pedraza

Department of Pathology and Kaplan Cancer Center and Department of Biochemistry, SUNY at Stony Brook, New York University Medical Center, NY 11796, USA

Received 18 March 1985; Revised and Accepted 14 May 1985

ABSTRACT

A cDNA clone was isolated for *Artemia salina* protein HD40, a component of heterogeneous nuclear ribonucleoproteins. Enriched *Artemia* 15S poly(A)⁺ RNA was used as a template and double-stranded cDNA sequences were inserted into the *Pst* I restriction endonuclease site of *E. coli* plasmid pBR322. Recombinant colonies were analyzed by positive hybrid selection of poly(A)⁺ RNA that directs the synthesis of protein HD40 in an *in vitro* assay. *In vitro* translation of the mRNA selected by anti-HD40 antibodies and that comigrates with authentic HD40 on gel electrophoresis. Partial proteolysis of protein HD40 and the *in vitro* translated product selected by clone 87HD produces the same peptide patterns. The size of the cloned insert is about 820 bp. The length of HD40 mRNA as determined by Northern blot analysis, is about 1500 nucleotides. Southern blot analysis performed with DNA of different species (plant, avian, mammal) shows cross-hybridizing bands when probed with clone 87HD DNA suggesting that the HD40 gene is evolutionarily conserved.

INTRODUCTION

In eukaryotic cells, mRNAs and their nuclear precursors hnRNAs, are complexed with proteins giving rise to ribonucleoprotein particles (RNPs). The elucidation of the role of the proteins which bind RNA is essential for understanding the cellular processes involving hnRNA and mRNA (1). In electron micrographs of transcriptionally active chromatin, hnRNP can be seen as nucleoprotein fibrils with 20 nm beads spaced along their length (2-4). The individual hnRNP beads can be recovered from purified nuclei by extraction with isotonic buffers at pH 8.0-9.0 as monoparticles that sediment at 30-40S. The particles are about 20 nm in diameter, contain 8-10S fragments of rapidly labeled RNA and a number of proteins that comprise about 80-85% of the particle mass. A substantial fraction of the protein mass of hnRNP consists of a group of basic proteins (pI = 8.0-9.0) with molecular weights between 30,000 and 45,000 characterized by similar

ACKNOWLEDGEMENTS

The authors thank Dr. E. L. Rothblum for providing the cloned rat rDNA sequences, Dr. A. H. Cavanough for providing the p1798 cell extract, and Ms. Betty Blum for preparation of the manuscript. This work was supported by National Institutes of Health Grants CA22394 to E.A.F. and 0128298 to H.R.S. and by American Cancer Society Grant 518-107 to H.R.S. H.R.S. is a recipient of National Cancer Institute Research Career Development Award CA00897.

REFERENCES

1. Thompson, E.A. (1980) *Mol. Cell. Endocrin.* 17, 95-102.
2. Cavanough, A.H. and Thompson, E.A. (1983) *J. Biol. Chem.* 258, 9768-9773.
3. Cavanough, A.H., Gietl, F. K., Lachert, R.P. and Thompson, E.A. (1984) *Proc. Natl. Acad. Sci. USA* 81, 718-721.
4. Smith, M.H., Reece, A.E. and Huang, R.C.C. (1978) *Cell* 15, 615-626.
5. Stallcup, H.R. and Washington, L.D. (1983) *J. Biol. Chem.* 258, 2802-2807.
6. Ringold, G., Lasfargnes, E.Y., Bishop, J.H. and Varmus, H.E. (1975) *Virology* 65, 135-147.
7. Ringold, G.H., Shank, P.R., Varmus, H.E., Ring, J. and Yamamoto, K.R. (1979) *Proc. Natl. Acad. Sci. USA* 76, 665-669.
8. Ardeman, R. (1979) *Gene* 7, 83-96.
9. Rothblum, E.L., Reddy, K. and Cassidy, B. (1982) *Nucleic Acids Res.* 10, 7349-7362.
10. Prescott, R.S., Washington, L.D. and Stallcup, H.R. (1984) *J. Virol.* 50, 60-65.
11. Wool, K.H., Bowman, L.H. and Thompson, E.A. (1984) *Mol. Cell. Biol.* 4, 822-828.
12. Mantatis, T., Jeffrey, A. and van de Sande, H. (1975) *Biochemistry* 14, 3787-3794.
13. Miller, K. and Solner-Webb, B. (1981) *Cell* 27, 165-174.
14. Wool, K.H. and Thompson, E.A. (1984) *Mol. Cell. Endocrinol.*, in press.
15. Sun, J. Y.-C., Johnson, E.H. and Allfrey, V.G. (1970) *Biochemistry* 18, 4572-4580.
16. Hipskind, R.A. and Reeder, R.H. (1980) *J. Biol. Chem.* 255, 7896-7906.

Nucleic acid sequence and chromosome assignment of a wheat storage protein gene

Ohm D. Anderson, James C. Lile, Mark Farnsworth and Frank C. O'Keefe

Food Protein Research Unit, Agricultural Research Service, U.S. Department of Agriculture, Western Regional Research Center, Albany, CA 91710, USA

Received 30 May 1984; Revised and Accepted 28 September 1984

ABSTRACT

A cloned gliadin gene was isolated from a wheat genomic library, and 2.4 kb of its primary sequence determined. The gene, α -1Y, was found by Southern analysis to be located on chromosome 6A, and its derived amino acid sequence identifies it as a member of the A-gliadin subgroup of α -gladins located on the short arm of that chromosome. α -1Y is apparently functional, and contains consensus TATA and CAAT boxes, and polyadenylation signals. This gliadin gene has no introns, and its noncoding flanking regions contain several short repeats and inverted sequences. The gene is contained in a 6.2 kb EcoRI genomic fragment whose apparent copy number varies in different wheat cultivars.

INTRODUCTION

The protein nutritional quality and unique rheological properties (dough-forming abilities) of wheat flour are determined largely by its principal storage proteins, the gladiins and glutenins (1). The gladiins are monomeric proteins of 30,000 - 78,000 molecular weight, and are characterized by low electrostatic charge density, poor solubility in dilute salt solutions, and good solubility in alcohol:water mixtures. They comprise a multigene family which has evolved by gene duplication and divergence from ancestral genes (1). They have been historically assigned to α , β , γ , and ω classes based on electrophoretic mobility, and more than 40 gliadin components can be detected by two-dimensional gel electrophoresis (2,3). The complete α -gliadin class, and most components of the β class are coded at loci located on the short arms of group 6 chromosomes of wheat; the complete ω class and most components of the γ class are coded at loci on the short arms of group 1 chromosomes (3,4). There is close linkage among the genes at each locus, and in intervallet crosses, they are inherited largely as nonrecombinant groups (3,4,5). The analysis of subfamily structure at each gliadin locus is incomplete because insufficient sequence and hybridization information is available, but at least one such

grouping has been recognized on the basis of the specific aggregation properties of its gene products, which are termed A-gliadins (1). This subfamily is coded at the 6A locus, has a mobility, and based on two-dimensional gel electrophoresis, contains at least 7 members (6).

Gliadin genes are expressed in the seed endosperm, under developmental control, probably at the level of transcription (7). A full complement of gene products is detectable at 6-9 days after fertilization, suggesting that the genes are coordinately activated (7,8, Greene unpublished).

Gliadin biosynthesis occurs in association with membranes (9), directed by long-lived mRNAs (7), and the presence of an *in*-terminal leader sequence has been confirmed (10,11). Sequence analysis of gliadin proteins and cloned gliadin cDNAs have yielded information on the coding regions of some members of this gene family (11,12), but no genomic sequences have been reported.

The evidence for close genetic linkage and coordinated expression of gliadin genes is consistent with a physical clustering in the wheat genome, and with the presence of similar control sequences in the genes. In order to investigate these facets of gene control further, we are pursuing a study of the gliadin loci in the wheat genome. The present report describes the isolation and structural analysis of a cloned gliadin gene coded at the 6A locus.

MATERIALS AND METHODS

Materials

Restriction enzymes were from Bethesda Research Labs, New England Biolabs and P-L Biochemicals. T4 ligase, DNA polymerase I, X-tal, Protase K, and acrylamide were from Bethesda Research Labs. Nitrocellulose was from Schleicher & Schell. Sequencing reaction mixtures, and DNA polymerase I Klenow fragment were from Bethesda Research Labs, and P-L Biochemicals. Hybridization primers and probe primers were from P-L Biochemicals. The X-ray film used was XAR-5 from Kodak. 32 P-dATP, dCTP, TTP and dGTP (>4000 Ci/mmol), 35 S-dATP (>1000 Ci/mmol), and Gene-Screen Plus hybridization membrane were from New England Nuclear. Low-melting agarose was from FMC. Zeta-Probe membrane was from Bio-Rad.

Isolation of gliadin genomic clones

Gliadin genomic sequences were isolated from a wheat (*Triticum aestivum*, cultivar Yamhill) library (13) constructed in Charon 32 (14) using DH1 (15) as host. Similar clones have been isolated from a cultivar Cheyenne library constructed by us (unpublished) in the vector Sep6-Lac5 (F. Meyerowitz, unpublished). Screening of gliadin clones was according to the methods of

Benton and Davis (16). The probes for all library screenings were restriction fragments of the gliadin cDNA clone pG-A10 (11). Plasmid subcloning of lambda inserts was accomplished by ligating an *Eco*RI digest of cloned DNA with *P*coRI restricted KVL7 DNA (described in (17)), or plasmid pUC8 (23). The Yamhill clone Yam-2 yielded the subclone pYAZ-28 (in KVL7), and the Cheyenne clone Chey-5 was the source of DNA for the subclone pChey-56 (in pUC8).

Analysis of gliadin clone YAM-2

M13 phage subcloning was performed by ligating fragments of the 6.2 kb (Figs 2 and 4) insert digested with four-base recognition restriction enzymes (Alu I, Hae III, Ksa I), with *Sma* I restricted vectors, or *Sau* 3A digested insert and *Bam* HI restricted vectors mp8-11 (18) to yield four sets of subclones. Coding region subclones were identified using a pT0-A10 probe. In some cases a sequenced clone was used to make a hybridization probe to isolate an overlapping sequence from a different subset of clones.

Sequencing reactions were by the di-deoxy-procedure of Sanger et al. (18). Hybridization probe was prepared as described by Hu and Messing (19). The conditions for both reactions were those suggested by P-L Biochemicals and Bethesda Research Labs.

Blots were performed as described by Southern (20), using nitrocellulose or nylon membranes under the following conditions: blots containing genomic and clone DNA were prehybridized for 48 hours and 16 hours, respectively, at 68°C in 1 M NaCl, 50 mM Tris 7.5, 5 mM EDTA, 200 μ g/ml sheared denatured salmon sperm DNA. Labelled gliadin DNA was added to fresh prehybridization buffer, and the blots incubated at 68°C for 24 hours (clone DNA) or 72 hours (genomic DNA). Blots were then washed once each at 68°C with 5 mM EDTA, 0.1% SDS plus the following: 2x SSC, 0.5% SSC, 0.1% SSC. Nick-translation of DNA fragments was according to Rigby et al. (21).

Stabilization energies of potential secondary structures were estimated according to the rules given by Tinoco et al. (22).

DNA Isolation

Single-stranded M13 DNA was prepared according to Messing and Vieira (23). M13 double-stranded DNA and plasmid DNA was isolated by the alkaline-SDS method of Birnboim and Doly (24) as described by Maniatis et al., (25). When necessary, the supercoiled DNA was further purified using CsCl equilibrium gradients or hydroxylapatite.

Wheat nuclei were prepared by modifications of the procedure of Iathe and Quatrano (26) using ethidium bromide as suggested by Kisilev and Mubinstein (27). DNA was isolated from nuclei by the Proteinase-K method



Figure 1. Screening of a wheat genomic library. The wheat lambda library was screened by the method of Benton and Davis (16). After the plaques were transferred to nitrocellulose filters, they were probed as described in Materials. Filters are to scale, and the plaques in B and C are the same size. A) 10,000 pfu of the total library on a 150 cm plate. Most filters contained only 1-3 detectable signals. B) 20 pfu of the 3rd plating of signal b on an 88 cm plate. C) 24 pfu of the 3rd plating of signal c on an 88 cm plate.

of Blin and Stafford (28). The isolation is described in more detail in Ulits et al. (in preparation).

Specific DNA fragments were isolated from low-melting agarose as described by Weislander (29). DNA ligations, cell transformations, lambda growth and lambda DNA isolation were all performed by the procedures described in Maniatis et al. (25).

RESULTS

In our initial screen, approximately 600,000 plaques from a wheat library (cultivar Yamhill) were probed with the labelled gliadin cDNA clone pTo-A10 (11). Figure 1A is an autoradiogram from a plate showing several positive clones displaying varying signal intensity. From 120 such plaques, 20 were carried through two additional cycles of purification to isolate single clones (Figure 1 B & C). These further cycles established that the different signal intensities were not due to plaque size, but likely

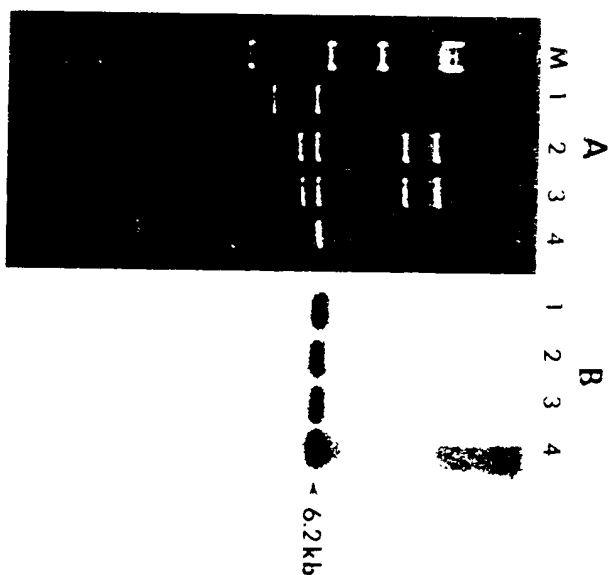


Figure 2. EcoRI restriction enzyme analysis of gliadin clones. DNA electrophoresed on a 1% agarose gel, stained with ethidium bromide (A) and probed with the [32 P]-labelled gliadin cDNA pTo-A10 (B). M: Hind III digest of lambda DNA. Lane 1, plasmid pYAZ-28 (6.2 kb fragment of YAM-2 subcloned into the plasmid pV11A7); Lane 2, lambda YAM-2; Lane 3, lambda Chey-5; Lane 4, plasmid pCIR-5b (6.2 kb fragment of Chey-5 subcloned into plasmid pUCB).

due to different degrees of homology of each clone with the cDNA gliadin probe. This result would be expected since the gliadins are a multigene family of evolutionarily related, but distinct members (1,12).

From 12 genomic clones giving strong signals to the gliadin cDNA probe, one of the strongest, YAM-2, was chosen for further analysis. When YAM-2 DNA was isolated and subjected to EcoRI restriction, the wheat insert yielded fragments of 5.5 and 6.2 kb (Figure 2A, lane 2), clearly separated from the lambda arms of approximately 11 kb and 19 kb. Only the 6.2 kb fragment hybridized with the gliadin cDNA probe. This fragment was subcloned into plasmid pV11A7 for further analysis (lane 1). A clone (Chey-5) isolated from the Cheyenne library is shown in lanes 3 & 4 for comparison.

A partial restriction map of the gliadin related 6.2 kb insert from YAM-2 is shown in Figure 3. The gliadin related sequence is approximately centered within the EcoRI fragment, between two Not I sites 1 kb apart. The 6.2 kb

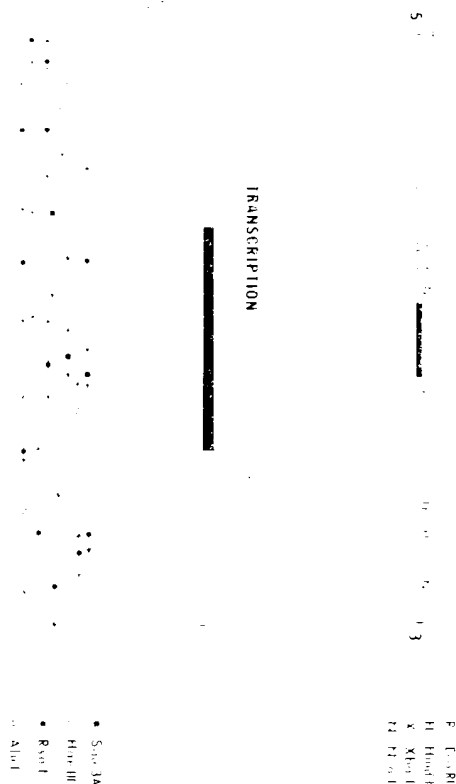


Figure 3. Partial restriction map of the YAM-2 6.2 kb insert. The map of the 6.2 kb fragment of YAM-2 into pYAZ-28 was determined for the indicated enzymes. The sequenced portion is shown expanded. Arrows below the map indicate the specific sequences determined with M13 subclones.

insert was restricted with four-base recognition restriction enzymes and the resultant fragments subcloned into M13 as described in Materials and the central portion of the 6.2 kb fragment sequenced as shown in Figure 3. Overlapping clones were assembled into the final sequence given in Figure 4.

Position 41 of the coding sequence of the YAM-2 gliadin was assigned because it is the only potential initiator codon for an open reading frame. There is no indication of introns interrupting the coding region. The nucleotide sequence codes for a protein of 286 amino acids and a molecular weight of 32,912. The amino acid sequence derived from this sequence is 97% homologous to the Arg141in protein amino acid sequence determined by Kasarda et al. (11). This confirms its identity as a member of the A-gliadin subfamily of gliadin storage protein genes. In addition, the YAM-2 sequence shares 93% and 96% nucleotide homology with the gliadin cDNA clones pG1A-42 and pTO-A10, respectively (11). A characteristic of these gliadins is the presence of two polyglutamine regions, of which the 5' polyglutamine region seems more conserved in length than the 3' one. In YAM-2, for example, the 5' polyglutamine has the sequence (CAG)₉-(CAA)₉, compared to (CAG)₅-(CAA)₁₂ in the gliadin cDNA clone pTO-A10. The nonrandom distribution of CAG and CAA may indicate mechanisms controlling codon usage within the region.

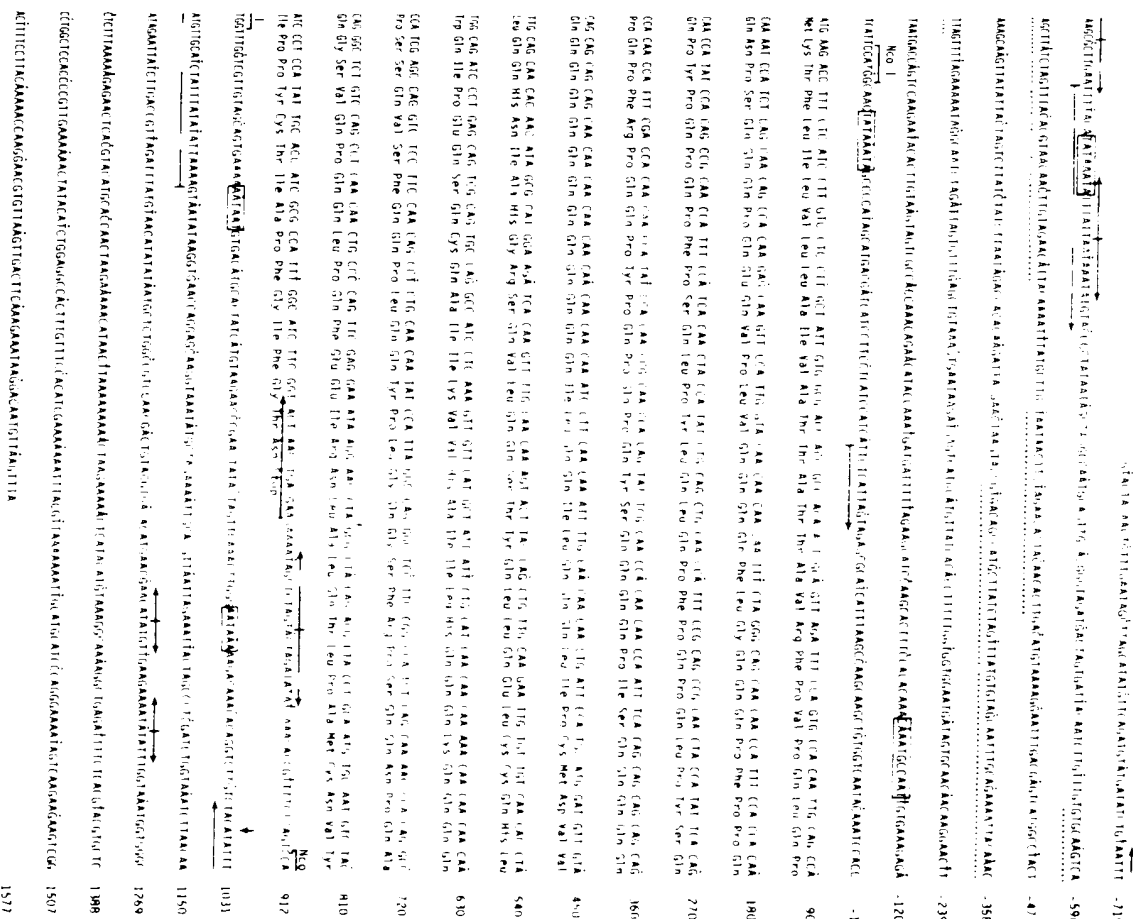


Figure 4. Sequence of a 2346 base region of the YAM-2 6.2 kb insert. The primary sequence is shown along with the translation of the open reading frame. Putative control elements are boxed. The dotted lines indicate 80% homologous repeated segments of the 5' flanking region. Underlined sequences are direct inversions. Underlined sequences found in the noncoding regions indicate inversions of sequences found in the opposite flanking region. An arrow points to the polyadenylation site of cDNA clones (11).

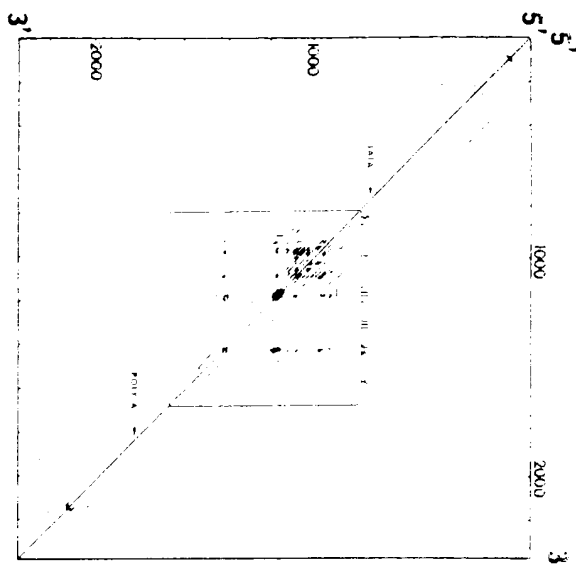


Figure 5. Homology matrix of the 2346 base sequence with itself. A homology matrix was plotted of the entire sequenced portion of YAM-2. A homology criterion of 14 bases out of 20 was used. The coding region is boxed and the domains of the Arg-lactin protein (11) are labeled. The presumptive 'TATA' and polyadenylation sites are indicated.

The 3' noncoding region of most messengers contains a putative polyadenylation signal related to AATAAA (30). Two such sites seem to be common in plant genes reported thus far (31, 32). These sequences have been shown to be necessary for proper polyadenylation of mRNAs (33). The 3' noncoding region of the YAM-2 sequence is 98% homologous to the 3' ends of two gliadin cDNAs reported by Kasarda et al. (11). The 3' region of all 3 sequences contain 2 potential polyadenylation signals, centered, in α -LY at +941 (AATAAT) and +998 (AATAAA).

The 5' end of the coding region of α -LY is established by the potential nonsense codon at -70 followed in frame at +1 by the only start codon (ATG) allowing correct reading of the following gliadin sequence. Nearby, in the 5' upstream sequence, are several sequences related to presumptive control elements discussed by Breathnach and Chambon (34). A 'TATA' sequence of TATAAAT, matching the consensus sequence of TATAAAT is found at position -104 (Figure 4 and Figure 7A). Thirty-seven bases further upstream from the 'TATA' at -141 is the sequence CAAATGCCAAT which contains two potential 'CAT' like elements.

In order to examine the internal sequence homologies within YAM-2, the sequenced portion was analyzed via homology matrix. Homologies from 60-90% and windows from 20 to 50 bases were used, with the 70% homology at a 20 base window shown in Figure 3 showing the main, consistent features of the analysis. Within the coding sequence, the five domains of the Arg-lactin primary sequence (11) and a signal sequence are discernible with the following characteristics: 5. The leader sequence coding for a 20 amino acid signal peptide with little external homology. 1. A 300 base region with extensive internal and little external homology. 2. The first polypeptide region. 3. A 200 base fragment with limited internal homologies. 4. The second polypeptide region. 5. A 200 bases 3'-terminal sequence with some internal homology in its 5' portion.

The matrix also points to several short homologies in the flanking regions. The 5' noncoding sequence contains a 300 base region from about -600 to -310 with several internal homologies, the highest of which is a 56 base repeat of 80% homology starting at -589 and -395. A third sequence, of 28 bases and starting at -339, shares as much as 82% homology with the first two sequences. In addition, there are two sequences of 31 bases sharing 77% homology (at -511 and -321). The 3' flanking region contained no significant external homologies, and short (10-15 bases) internal homologies mainly in its more distal sequence from the coding region.

Three potentially significant inverted repeat structures occur in the noncoding sequences (see Figure 4). The first is centered at -690. It is comprised of two contiguous nine base direct inversions, one with eight of nine bases matching, and the other of nine perfect matches. This region also has the potential to form a cruciform-like structure with a 3' sequence at +1023 as shown in Figure 7C. A second interesting sequence occurs at +1230. Here, within 26 bases, are two 5-base direct inversions, a ten base perfect repeat of the sequence following the coding termination codon, and a nine base inversion of the sequence beginning at -745. Finally, at +852 is a 13 base sequence, including the termination codon, which is an imperfect inversion of the sequence at -59, which includes one of the two potential mRNA start sites. This pair is of interest because it involves two important locations, the beginning of transcription and the termination of translation. What functional significance this relationship has is at present unknown, but we note that several other plant genes have short inversions involving the termination codon and a sequence between the TATA box and the initiator codon (31, 32, 35, 36) although the

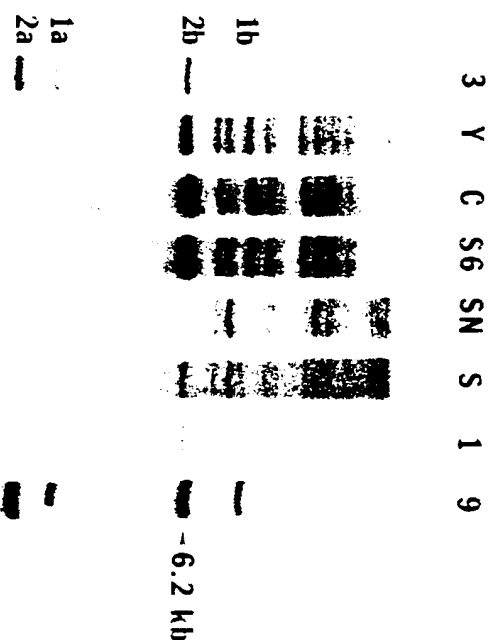


Figure 6. Southern analysis of wheat genomic DNA with a gliadin probe. Fifteen micrograms of total nuclear DNA of the indicated wheat cultivars was digested with *Eco*RI and electrophoresed on a 0.7% agarose gel. The gel was blotted and probed, as described, with the 32p labelled 1.1 kb *Not* I fragment of YAH-2 containing the entire coding region of the gene. Y, Yamhill, G, Cheyenne; S6, Chinese Spring with a Cheyenne 6A substitution; SN, Chinese Spring nulli-6A-tetra-6B; S, Chinese Spring. Control bands; 2a, 3.0 Hind III fragment of YAH-2 containing the entire coding region of the gliadin gene *wt*-1; 2b, 6.2 kb *Eco*RI fragment of YAH-2; 1a and 1b, derived from the clone YAH-1.

degree of homology is not always as great as with the present sequence.

Southern blot analysis was employed in determinations of copy numbers and chromosomal locations of the gliadin genes. Blots of total Yamhill and Cheyenne DNA probed with the *hco* 1 - *hco* 1 coding region fragment of α -IY revealed a series of hybridizing bands from an intense 6.2 kb band to fainter, higher molecular weight bands of up to 20 kb (Figure 6). Similar blot patterns were obtained using probes derived entirely from within the coding region of gliadin cDNA clone PTO-A10 (data not shown), indicating that the pattern represents gliadin-related gene fragments. In addition, these patterns have been consistently observed under digestion conditions in which both time and enzyme/DNA ratios were varied, indicating that they represent limit digestions. Yamhill (Y), Cheyenne (C) and Chinese Spring (S) all contain

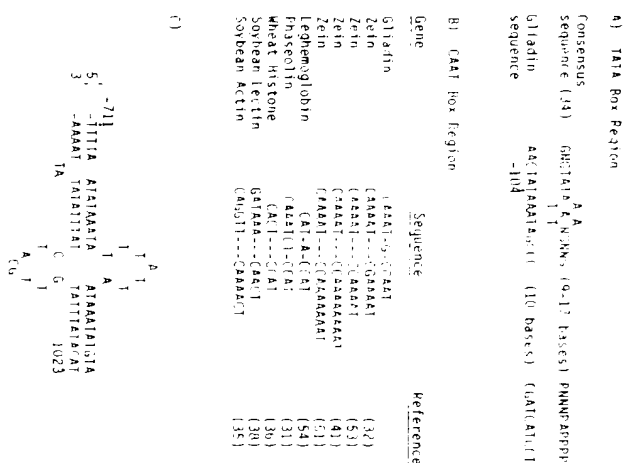


Figure 7. Specific sequences within the 2346 base fragment. The putative 'TAAAT' of the glutadin gene alpha-17 is compared to the consensus sequence of Breathamch and Chlambo (34). N: purine or pyrimidine; P: pyrimidine. B) The 'CAAT' sequence at ~131 of glutadin gene- α 17 is compared to similar reported sequences in other plant genes. C) The secondary structure that could potentially form between the sequences at ~711 and +1022.

the 6.2 kb band, though with different intensities, suggesting different copy numbers of 6.2 kb gliadin sequences. Copy number estimates were based on reconstructions using standard bands 2a & 2b derived from genomic clone YAM-2 mixed with a blank hybridizing background of sea urchin DNA. YAM-2 has been shown to be a member of the 6.2 kb gliadin gene group. Partial sequencing (unpublished) has established that YAM-1, which contains a gliadin gene within a 7.8 kb EcoRI fragment, belongs to a subfamily of gliadins closely related to, but distinct from the subfamily of YAM-2. The copy number estimates for the 6.2 kb band indicate 1-3 copies in Chinese Spring, 15-20 copies in Cheyenne, and an intermediate number in Yamhill. These estimates assume close homology with the gene *g1-1* (isolated as clone YAM-2) whose coding region was used to probe the blot. EcoRI fragments of lower homology (such as YAM-1) would yield a lower apparent copy number. This appears to be the case with sequence at 6.2 kb in Chinese Spring multi-6A-tetra-6B whose intensity is equivalent to about 0.1 copy per genome. When the

junctions (35) is present in three positions in the coding sequence of this Argliadin gene; 670-674, 681-685, 721-725 (Fig. 4), no actual introns are indicated. The gliadin genes are considered to have resulted, in part, from duplications of shorter ancestral sequences (6,11,12), but the present information suggests either that the evolution of Argliadin genes did not involve intron/exon structures, or that such structures were eliminated during the evolution (see 45,46,47).

The α -IY sequence contains all of the recognized consensus control regions. The sequence GAAATGCGAT located 37 bases upstream from the 'TATA' box is particularly interesting in that such 'CAAT-CAAT' structures are present in several, but not all, plant genes thus far reported (Figure 7b). Further sequences from a wider variety of species and genes are needed to establish the distribution and variability of this region. The precise functional role of specific portions of this region is yet to be determined, particularly in light of conflicting reports as to its functional importance in *in vitro* mutated genes (38,39,50).

Langridge and Felix (51) have reported two promoter regions in a zein gene, yielding transcripts of two lengths. An examination of the α -IY sequence shows a 'TATA'-rich region at approximately -730 bp, in addition to the TATA at -104. Interestingly, the more distal 'TATA' region includes a direct 9 base inverted repeat, similar to the zein reported by Langridge and Felix, but lacking the internal loop of the zein. In addition, Langridge and Felix found a 15 base direct repeat. α -IY, instead, contains a larger region with several 70-80% homologies which may be the remnants of ancient duplications.

This distal 'TATA' is part of the first region of inversions mentioned earlier (centered at -690). In addition, the sequence from the 3' part of this inversion begins at +1023, the polyadenylation site in 2 cDNA clones (11). The potential significance of these sequences must await further analysis of the gliadin multigene family and transcriptional studies to delineate those sequences necessary for gene activity.

We have also isolated gliadin clones from a wheat genomic library (cultivar Cheyenne) constructed (Anderson et al., in preparation) in vector lambda Sep6-lac5 (E. Meyerowitz; unpublished, see (25)). The wheat cultivars Yamhill and Cheyenne show similar patterns in the A-gliadin regions in 2-D PAGE of seed proteins (unpublished). One of the Cheyenne clones, Chey-5, is similar to Yam-2 and is also shown in Figure 2. Copy number estimates made from Southern blots of wheat cultivars Yamhill and the related Cheyenne (Figure 3, and (52)) indicate that there are 1-20 copies of 6.2 kb EcoRI

gliadin fragments in the wheat A genome, with the exact number dependent on each specific wheat cultivar. The results of the Southern analysis shown on Figure 6 indicates potential changes in the 6.2 kb gliadin sequences, either by expansion or duplication of the total number of members of this group. Further study will delineate if this group represents a duplication and possible divergence within a contiguous locus on the 6A chromosome. Support for a group of similar, but distinct genes comes from a further restriction analysis (unpublished) of Chey-5 and Yam-2 which indicates similar, but not identical restriction patterns for the two clones.

These questions of duplications and divergence will be resolved as we expand our study into the rest of the Argliadin gene sub-family.

ACKNOWLEDGMENTS

We wish to thank M. Murray and J. Slightom (Agrigenetics) for sending us the wheat genome library. The Chinese Spring nullisomic-6A-tetra-6R was generously provided by E. K. Sears (37).

Permanent address: Laboratoire de Technologie Alimentaire, Institut National de la Recherche Agronomique, 9 Place Viala, 34060 Montpellier, France.

REFERENCES

1. Kasarda, D. D., Bernardin, J. E., and Nimmo, C. C. (1976) Wheat Proteins, In *Advances in Cereal Science and Technology*, Vol. 1, Y. Pomeroy, ed., pp. 158-236, American Association of Cereal Chemists, St. Paul, Minnesota.
2. Wrigley, C. W., and Sheppard, K. W. (1973) *Ann. N.Y. Acad. Sci.* 209, 154-162.
3. Meham, D. K., Kasarda, D. D., and Onalset, C. O. (1978) *Biochem. Genet.* 16, 831-853.
4. Payne, P. I., Jackson, E. A., Holt, L. M., and Law, C. N. (1984) *Theor. Appl. Genet.* 67, 235-243.
5. Sozinov, A. A., and Popereleva, E. A. (1980) *Ann. Technol. Agric.* 29, 229-245.
6. Kasarda, D. D. (1980) *Ann. Technol. Agric.* 29, 151-173.
7. Greene, F. C. (1983) *Plant Physiol.* 71, 40-46.
8. Meham, D. K., Follington, J. G., and Greene, F. C. (1981) *J. Sci. Ed. Agric.* 32, 773-780.
9. Greene, F. C. (1981) *Plant Physiol.* 68, 778-783.
10. Okita, T. W., and Greene, F. C. (1982) *Plant Physiol.* 69, 834-839.
11. Kasarda, D. D., Okita, T. W., Bernardin, J. E., Baeker, P. A., Nimmo, C. C., Lew, E. J.-L., Dietler, M. D., and Greene, F. C. (1984) *Proc. Natl. Acad. Sci. U.S.A.* In press.
12. Bartsels, D., and Thompson, R. D. (1983) *Mol. Acids Res.* 11, 2961-2977.
13. Murray, M. G., Kennard, W. C., Drong, R. F., and Slightom, J. L. (1984) *Gene* (in press).
14. Loenen, W. A., and Blatner, F. (1983) *Gene* 26: 171-179.
15. Ibanhan, D. (1983) *J. Mol. Biol.* 166: 557-580.
16. Benton, W. D., and Davis, K. W. (1977) *Science* 196: 180-182.
17. Lynn, D. A., Angerer, L. M., Bruskun, A. M., Kleinf, W. H., and Angerer, R. C. (1983) *Proc. Natl. Acad. Sci. USA* 80: 2656-2660.
18. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* 74: 5463-5467.

19. Ito, N.-I., and I. Tessing, J. (1982) *Gene* **17**, 271-278.
20. Southern, E. (1975) *J. Mol. Biol.* **98**, 503-517.
21. Rigby, P. W. J., Dieckmann, T., Rhodes, C., and Berg, P. (1977) *J. Mol. Biol.* **113**, 237-251.
22. Hsu, H.-I., Royer, J. H., Dangler, B., Levine, H. D., Mullenbeck, O. E., Crothers, D. H., and Gaitly, J. (1973) *Nature New Biology* **236**, 40-41.
23. Messing, J., and Vieira, J. (1982) *Gene* **19**, 269-276.
24. Birnboim, H. C., and Doly, J. (1979) *Nucleic Acids Res.* **7**, 1513-1523.
25. Maniatis, T., Fritschy, E. F., and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Press, NY.
26. Lathrop, D. S., and Orian, R. (1980) *Plant Physiol.* **65**, 305-308.
27. Kistley, L., and Kohnstien, L. (1980) *Plant Physiol.* **66**, 1140-1143.
28. Bilo, K., and Starck, D. W. (1976) *Nucleic Acids Res.* **3**, 2303-2308.
29. Weistlander, U. (1979) *Ann. Biochem.* **93**, 305-309.
30. Proudfoot, N. J., and Brownlee, G. (1976) *Nature* **263**, 211-214.
31. Slightfoot, T. L., Sun, S. H., and Hall, F. C. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 1697-1701.
32. Pederson, F., Lawrence, L., Wilson, D. K., Sheldon, E., and Larkins, B. A. (1982) *Cell* **29**, 1015-1026.
33. Fitzgerald, M., and Shuh, L. (1981) *Cell* **24**, 251-260.
34. Berthiauch, K., and Chabot, P. (1981) *Ann. Rev. Biochem.* **50**, 349-383.
35. Shuh, D. H., Hightower, K. C., and Heaghter, K. B. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 1022-1026.
36. Tabata, T., Sasaki, K., and Iwabuchi, H. (1983) *Nucleic Acids Res.* **11**, 5865-5875.
37. Sears, E. K. (1975) *The Ancestry of Common Wheat. Res. Bull.* (Missouri Agric. Expt. Sta.) Vol. 572.
38. Volkin, I. O., Phillips, P. K., and Goldberger, K. B. (1983) *Cell* **34**, 1023-1031.
39. Schuler, M. A., Schmitt, E. S., and Beachy, K. N. (1982) *Nucleic Acids Res.* **10**, 8225-8245.
40. Montell, C., Fischer, E. F., Caruthers, M. H., and Beck, A. J. (1983) *Nature* **305**, 600-605.
41. Spena, A., Viotti, A., and Pirodda, V. (1982) *EMBO Journal* **1**, 1589-1594.
42. Pearson, W. J., Dennis, E. S., Ellis, J., Flanagan, E. J., Gerlach, W. L., Llewellyn, D., and Sachs, M. M. (1984) *J. Cellular Biochem. Abstracts* **88**, 15.
43. Shuh, D. H., Hightower, K. C., and Heaghter, K. B. (1983) *J. Mol. Appl. Genet.* **2**, 111-126.
44. Fischer, K. L., and Goldberger, K. B. (1982) *Cell* **29**, 651-660.
45. Zakut, R., Shani, M., Glivol, D., Neuman, S., Yaffe, D., and Nudel, U. (1983) *Nature* **298**, 857-859.
46. Go, M. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 1964-1968.
47. Efferman, F. A., Yound, P. K., Scott, K. W., and Tilghman, S. M. (1981) *Nature* **294**, 713-718.
48. McKnight, S. L., and Kingsbury, K. (1982) *Science* **217**, 316-324.
49. Grosveld, G. C., deBoer, E., Siewmaker, G. K., and Flavell, K. A. (1982) *Nature* **295**, 120-126.
50. Dicks, F., van Ooyen, A., Cochran, M. D., Dobbin, C., Kelsner, J., and Weissman, C. (1983) *Cell* **32**, 695-706.
51. Langridge, P., and Felt, G. (1983) *Cell* **34**, 1015-1022.
52. Iltis, I. G., Anderson, O. D., Okita, T. W., and Greene, F. C. (1983) *Plant Physiol.* **72**, (Suppl.), 2.
53. Spena, A., Viotti, A., and Pirodda, V. (1983) *J. Mol. Biol.* **169**, 799-811.
54. Brisson, N., and Verma, D. P. S. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 4055-4059.

tion of cyclic AMP receptor protein with the *dhfr* biosynthetic operon in *E. coli*

Dr. Hiden, Dr. Jan, Kenmore of London and Martin Hearnish

Journal of Biochemistry, State University of New York, Stony Brook, NY 11794, USA

and Dr. Alan 1984; Received and Accepted 15 October 1984

ABSTRACT

Dhfr and restriction site protection studies show that cAMP and its receptor protein (CRP) bind to the promoter of the *dhfr* operon at approximately position -34 to -82. This region contains sequences that are homologous to those found in other CRP-dependent promoters. In vitro transcription from the *dhfr* promoter was markedly increased by the addition of cAMP and CRP. This stimulation was not found when the *dhfr* template lacked a proposed CRP binding site. cAMP-CRP did not alter the extent of transcription termination within the *dhfr* leader suggesting that this regulatory system may be independent of the attenuation mechanism involved in negative control of this operon. The results of restriction enzyme site insertion studies and experiments with altered promoter fragments indicate that the mechanism for CRP stimulation of the *dhfr* operon may be similar to a recently proposed for *lac* (1).

INTRODUCTION

The *dhfr* operon of *Escherichia coli* K-12 contains the structural gene for dihydroxy acid synthase I, an enzyme required for the biosynthesis of alanine, valine and leucine. Regulation of this operon is complex, involving negative control by attenuation (2,3) and positive control by a number of factors (4,5) including cAMP-CRP (6,7). The participation of cAMP-CRP in the regulation of a biosynthetic operon is very unusual since this complex is normally involved in the regulation of degradative operons (7). A recent report suggests that this control of *dhfr* may reflect a need to increase dihydroxy acid synthase I when the flow of carbon, in the form of the substrates of the enzyme, is reduced (8). An examination of the DNA sequence of the *dhfr* promoter revealed structural features similar to the P-binding site consensus sequences proposed for other operons (2). It was therefore of interest to determine directly if CRP binds to the *dhfr* promoter and to further investigate the effect of cAMP-CRP on *in vitro* transcription in this operon. In addition, since this is the first operon shown to be controlled by attenuation and cAMP-CRP, it was feasible to investigate

ACKNOWLEDGMENTS

We would like to thank Mark Zall, C. Caroline Boyd, Colin May and Debbie Cool for helpful suggestions during this study. This work was supported in part by grants from the Medical Research Council of Canada (MA-7716) and the British Columbia Health Care Research Foundation (19082-1). MRL was supported by a studentship from the MRC of Canada.

REFERENCES

1. Jackson, C.M. and Newerson, Y. (1980) *Ann. Rev. Biochem.* **49**, 765-811.
2. Fujikawa, K., Chan, M.H., Loefer, M.E. and Davie, E.W. (1974) *Biochemistry* **13**, 5290-5299.
3. Hittory, P. and Esnouf, M.P. (1974) *Nature New Biol.* **252**, 90-92.
4. Rosenbly, J.S., Butler, D.L. and Rosenbly, R.D. (1975) *J. Biol. Chem.* **250**, 1607-1617.
5. Craves, C.B., Matus, T.W., Willingham, A.K. and Strauss, A.M. (1982) *J. Biol. Chem.* **257**, 13108-13114.
6. McGilivray, R.L.A. and Davie, E.W. *Biochemistry* (in press).
7. Dagen, S.L., Fritzsche, H.C. McGilivray, R.L.A. and Davie, E.W. (1983) *Biochemistry* **22**, 2087-2097.
8. Jager, H., de la Salle, H., Schreiber, F., Ballmaier, A., Kohli, V., Finkler, A., Fritzsche, H. and Loefer, J-P. (1983) *Nucleic Acids Res.* **11**, 2325-2335.
9. Kierich, K. and Davie, E.W. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 6461-6464.
10. Titani, K., Fujikawa, K., Feilfeld, D.L., Ericsson, L.H., Walsh, K.A. and Neurath, H. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 4082-4086.
11. Chacinas, G. and van de Sande, J.H. (1980) *Methods in Enzymol.* **65**, 75-85.
12. Hanahan, B. and Meselson, M. (1980) *Gene* **10**, 63-67.
13. Grunstein, M. and Hogness, D.S. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 3961-3965.
14. Maxam, A. and Gilbert, W. (1980) *Methods Enzymol.* **65**, 499-560.
15. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463-5467.
16. Messing, J. (1983) *Methods Enzymol.* **101**, 20-78.
17. Sanger, F. and Coulson, A.R. (1978) *FEBS Lett.* **87**, 107-110.
18. Barnes, W.M., Brown, M. and Son, P.H. (1983) *Methods Enzymol.* **101**, 98-122.
19. Enfield, D.L., Ericsson, L.H., Fujikawa, K., Walsh, K.A., Neurath, H. and Titani, K. (1980) *Biochemistry* **19**, 659-667.
20. Steiner, D.F., Quinn, P.S., Chan, S.J., Marsh, J. and Tager, H.S. (1980) *Ann. N.Y. Acad. Sci.* **343**, 1-16.
21. Schroeder, W.A., Shelton, J.B. and Shelton, J.R. (1969) *Arch. Biochem. Biophys.* **130**, 551-556.
22. Thibodeau, S.M., Palmiter, R.D. and Walsh, K.A. (1978) *J. Biol. Chem.* **253**, 9018-9023.
23. Strauss, A.W., Bennett, C.A., Donohue, A.M., Rokev, J.A., Boime, I. and Alberts, A.W. (1978) *J. Biol. Chem.* **253**, 6270-6274.
24. Gordon, L.L., Buletier, K.A., Sims, H.F., Fadelstein, C., Seann, A.M. and Strauss, A.W. (1983) *J. Biol. Chem.* **258**, 14054-14059.

The complete nucleotide sequence of a legumin gene from pea (*Pisum sativum* L.)

Gianley W. Joyce*, Ronald R. D. Joy, and H. Shusud and Donald Boulter

Department of Botany, University of Durham, South Road, Durham DH1 1TA, UK

Received 1 March 1984; Revised and Accepted 2 May 1984

ABSTRACT

One of several genes coding for the major pea storage protein, legumin, has been completely sequenced. The sequence covers the whole of the transcribed coding, flanking, and 5' and 3' untranslated sequences. The protein sequence starts with a signal peptide and is followed by the legumin polypeptide sequence of 4,481 and the polypeptide sequence of 4,481. Compared to other legumin sequences, the 5' and 3' untranslated sequences, encoded by this legumin gene, which are 1,000 and 1,000 bp, respectively, are relatively rich in the sulphur amino acids. The coding sequence is interrupted by three introns which show secondary structure typical of higher plant genes. The 5' end of the gene sequence contains a "TATA box", a "CAAT box" and a sequence showing some homology to an "ATA box". An extra sequence, identical to the normal polyubiquitin of an animal or the legumin message is seen in the 3' untranslated region. The structure of the gene and the possible significance of secondary structure in the normal RNA transcript in affecting the level of polyubiquitination are discussed.

INTRODUCTION

Legumin, one of the major storage proteins of pea (*Pisum sativum* L.), is a hexameric protein. Each of the six constituent monomers consist of a 40 kDa polypeptide chain and a 10 kDa polypeptide chain. The polypeptide chain is linked by disulphide bonds (1,2). These two polypeptides are generated by post-translational proteolysis of a single polypeptide chain of 100 kDa (3,4,5,6).

Legumin is found in the seeds of many leguminous (7) and non-leguminous plants. The important crop plants in this category (cycloper, navy, broad bean, chick pea) and (Avena, alfalfa, and rice) (8,9,10,11) have all been shown to contain proteins of similar subunit structures which show sequence homology to pea legumin and which are also linked by disulphide bonds (12,13,14,15). Legumin is important in leguminous crops such as peas, broad beans and soybeans because it is a major storage protein and source of sulphur amino acids in seed meals (16). Sulphur amino acid levels are the main nutritional limitation of legume seeds and the production of lines which

proposed by the structure which appears in of subunit amino acid rich leucine is given and discussed here.

The leucine group of proteins is therefore a candidate for in vitro or *in vivo* control by protein modification. Whilst it is known that heterogeneity at the level of the protein (11) and mRNA (13) is reflected by the presence of several genes (3), little is known about the structure of these genes except for what has been deduced from pea leucine cDNA sequences (14) and also from radiographic studies of soybean leucine genes (15).

The level of leucine gene structure may throw further light upon the control of seed protein synthesis, since there is good evidence that control is primarily at the transcriptional level in pea (16,17,20) and soybean (17,22). Several sequences thought to be involved in transcription have been identified in animal systems (see 23), but it has been suggested that plant genes may differ in some other respects (24,25). In order to obtain a better understanding of the leucine gene structure and control we have cloned several leucine genes from pea (3). Here we report the complete nucleotide sequence of one leucine gene, including the whole protein coding region, three introns and the 5' and 3' flanking sequences.

MATERIALS AND METHODS

Materials

Deoxyribonuclease I (DNase I) was obtained from Walthamton Biochemicals (Mullipore Ltd., London). Bovine alkaline phosphatase, endonuclease free DNA polymerase I and T4 polynucleotide kinase were from Boehringer Corporation Ltd. (London UK). Restriction endonucleases were from Boehringer Corporation Ltd. (Boehrle Research Laboratories (Cambridge UK) or New England Biolabs (or Laboratories Ltd., Bishop's Cleeve, UK). Digoxigenin-labeled triphosphates were from P-L Biochemicals Ltd., (Northampton UK) and γ -³²P ATP (50 Ci mmol⁻¹) was from Amersham International Ltd. (Amersham UK).

Genomic clones

Full details of cloning and isolation of λ leg 1 and other leucine genomic clones from gene libraries of Pisum sativum, cv 'Foltham First' are described elsewhere, as is the construction of the sub-clones pMIR21 and pMIR21 (16).

DNA sequencing

Double stranded 5' end-labeled DNA fragments were prepared as described

by Maxam and Gilbert (26) except that 3' protruding ends were labeled by incubating phage tails treated with γ -³²P ATP at 37°C with 100 μ Ci of γ -³²P ATP (30 Ci mmol⁻¹) and 50 units of polynucleotide kinase in 50 μ l of 0.1 M Tris-HCl pH 7.5, 10 mM MgCl₂, 1 mM dithiothreitol, 0.1 mM EDTA, 0.2 mM spermidine. The end-labeled fragments were sequenced by an improved nick-translation method: the 16 base-pair-kilobase method of Gilbert et al. (27). The only modifications were that end nucleotides from pMIR21 were used, phase 1 (1 and 2) was added to the reaction mixtures to give a 1:1 ratio of control labeling, and each base pair mixture was pooled with the corresponding forward mixture prior to hybridisation and analysis.

DNA sequences were analysed both by eye and by employing a standard computer programme (28). RNA secondary structure predictions were computed with the aid of a programme produced by Zuker and Stiegler (29). SI nuclease mapping

End labeled RNA fragments were prepared as above. Hybridisation and SI treatment were based on the methods of Faldutaro et al. (30) and Pedersen et al. (31). The amount of labeled RNA fragments prepared from 1 μ g of plasmid DNA (0.5 μ g) and 2 μ g of poly A⁺ RNA from developing pea cotyledons were used for each hybridisation together with 20 μ g Escherichia coli RNA carrier. Control reactions contained no poly A⁺ RNA. The nucleic acids were ethanol precipitated and resuspended in 20 μ l of 80% adjusted formamide, 0.1 M NaCl then incubated at 30.5°C for 1 hr. The samples were diluted with 200 μ l of ice-cold 50 mM NaCl, 30 mM Tris-HCl pH 7.5, 1 mM ZnSO₄ 7H₂O, adjusted containing 2500 units of SI nuclease and incubated at 37°C for 60 min. The samples were then chilled and ethanol precipitated (col) using the addition of a further 20 μ l of 0.1 M Tris-HCl, 0.1 M formamide, 10 mM NaOH, 1 mM EDTA, 0.1% xylene cyanol, 0.1% 10% pyridine blue were added and the samples were loaded onto a normal sequencing gel alongside four sequencing tracks prepared as above.

RESULTS AND DISCUSSION

Cloning and Sequencing

The cloning and isolation of several pea genomic fragments coding for leucine in lambda vectors is described elsewhere (16). A 1.6 kb EcoRI fragment inserted into the vector λ of WIS λ R (16). Based on results from restriction mapping and Southern blotting analysis using pea leucine cDNAs pMIR21 and pMIR21 (15), none of clones from λ leg 1

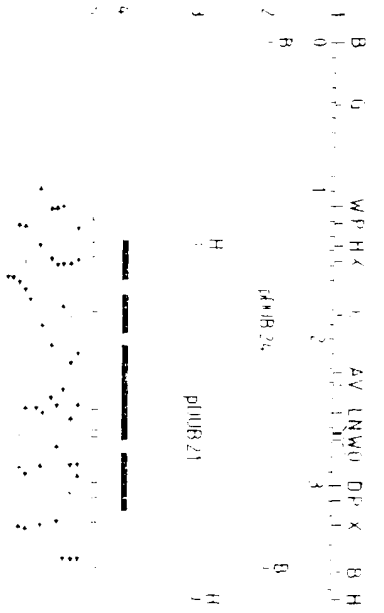


Figure 1. The protein structure of the pHUB21 and pHUB22 proteins. The figure shows the protein structure of the pHUB21 and pHUB22 proteins. The protein structure of the pHUB21 and pHUB22 proteins is shown in the figure. The protein structure of the pHUB21 and pHUB22 proteins is shown in the figure. The protein structure of the pHUB21 and pHUB22 proteins is shown in the figure.

The protein structure of the pHUB21 and pHUB22 proteins is shown in the figure. The protein structure of the pHUB21 and pHUB22 proteins is shown in the figure. The protein structure of the pHUB21 and pHUB22 proteins is shown in the figure. The protein structure of the pHUB21 and pHUB22 proteins is shown in the figure.

The protein structure of the pHUB21 and pHUB22 proteins is shown in the figure. The protein structure of the pHUB21 and pHUB22 proteins is shown in the figure. The protein structure of the pHUB21 and pHUB22 proteins is shown in the figure. The protein structure of the pHUB21 and pHUB22 proteins is shown in the figure.

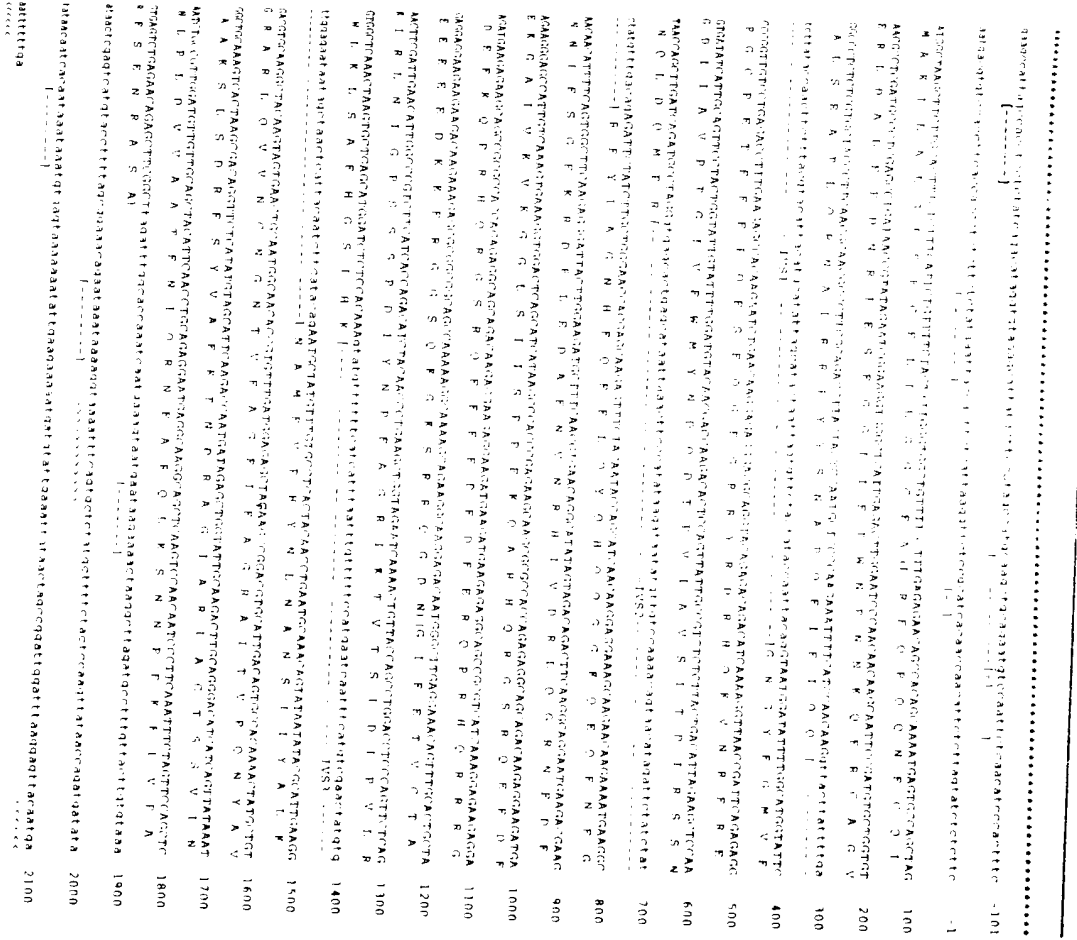


Figure 2. The sequence alignment of the pHUB21 and pHUB22 proteins. The alignment shows the amino acid sequences of the two proteins, with gaps indicated by dashes. The sequences are aligned from position 1 to 2000.

Proteinaseogenically from a reported different legumin gene (Leg 1) data not previously confirms that the reported present in more than one gene, but it is still not certain whether there is that the reports exist in gene.

Leg 1 and Proteinaseogen

The Leg 1 gene is present in the proteinaseogen from the gene sequence and is shown in Table 1. Both of the legumin genes have been identified previously from the Leg 1 gene in a previous and from several different cDNA clones (5,15) but this is the first complete sequence including the N-terminal region of the α-subunit from the Leg 1 gene. From the first identified sequence with the N-terminal cDNA sequence, it is now determined by the Leg 1 gene (13) shows a match of 100% of 25 amino acids, the two different Leg 1 genes account for 100% of the amino acids. In the gene, the Leg 1 gene with other seed proteins of Leg 1 (14,15) and other plant species (the Leg 1, 16, 17, 18, 19, 20, 21) there is a short peptide sequence rich in hydrophobic residues beyond the N-terminus of the mature protein, which is presumably a signal peptide (22). Unlike the other evidence for Leg 1 in the Leg 1 gene sequences (13, 14, 15, 16), the presence of signal peptides in Leg 1 protein precursor polypeptides has not been demonstrated previously by cDNA cloning and sequencing, though one in vitro synthesis study may be interpreted as showing such a signal peptide (20). Earlier reports have been demonstrated in legumin precursors from other legumes, namely Vicia faba (17) and Glycine max (20) synthesized by 'in vitro' protein synthesis systems.

In order to predict the accurate size and composition of the protein substrate produced by the Leg 1 gene, it is necessary to know the exact site of post-translational proteolysis. It was recently suggested by analogy with animal peptide prohormone processing that legumin might be post-translationally cleaved between the paired basic residues, five amino acids upstream from the N-terminus of the α-subunit, as well as adjacent to the N-terminus, leading to the removal of a five residue peptide (5). More recently, the C-terminal 38 residue of an α-subunit of legumin was isolated and identified by its N-terminal sequence and amino acid composition. This peptide was shown to extend to the aspartate residue adjacent to the N-terminus of the α-peptide (15). Carboxypeptidase A digestion has confirmed that aspartate is the C-terminal residue (d. citry pers. comm.). It is clear therefore, that the cleavage of at least some Leg 1 protein precursors occurs at a single site, in contrast to a report of the removal of a linking peptide between the A and B subunits of soybean legumin (23). This single

Table 1. Predicted amino acid composition of Leg 1 gene product

Amino acid	no. of residues	α-subunit no. of residues	β-subunit no. of residues	40% subunits no. of residues
Ala	12	(4.9)	11	(11.9)
Arg	6	(11.9)	13	(6.9)
Asn	20	(0.1)	1	(1.9)
Asp	13	(0.5)	1	(1.9)
Cys	3	(1.0)	2	(1.1)
Glu	33	(10.4)	1	(2.2)
Gln	43	(13.8)	9	(4.9)
Gly	27	(8.2)	10	(5.4)
His	7	(2.9)	3	(1.5)
Ile	14	(4.5)	9	(4.8)
Leu	15	(4.8)	21	(11.4)
Lys	13	(1.2)	4	(1.9)
Met	3	(1.0)	1	(0.5)
Phe	12	(3.9)	2	(0.2)
Pro	15	(4.8)	4	(1.9)
Ser	11	(4.5)	15	(5.1)
Thr	6	(1.9)	9	(4.9)
Tyr	2	(0.1)	1	(0.5)
Val	8	(2.6)	5	(2.2)
	11	(3.6)	11	(2.6)
			25	(5.0)

processing site in Leg 1 legumin resembles the internal cleavage points of zein in that it is adjacent to an aspartate residue preceded by an acidic residue (14,15).

There is considerable variation in the reported sulphur amino acid content of legumin from different Leg 1 strains (5). Since a single inbred strain of Leg 1 may contain a number of legumin genes (5) and several legumin protein types (14,15), the reported amino acid composition of the isolated legumin protein is undoubtedly a composite of several protein variants. As the low sulphur content of Leg 1 seeds is the main nutritional limitation, the isolation of genes coding for high sulphur legumin is an important first step toward crop improvement. The Leg 1 gene codes for a protein having 5 cysteine residues (1.01 mol %) and 4 methionine residues (0.94 mol %), which is at the upper end of the range of variation found between legumins of

LeuA gene homology

	EXON	INTRON	EXON
LeuA 1 (3-1)	A C C A A T C T T A C C T T A T T		
LeuA 1 (2-2)	C T C C C C C C T C C C C C C C C C		
LeuA 1 (3-3)	A C C A A A C C C T T A C C T T T T T		
P	C C C C C C C C C C C C C C C C C C		
2	A A C C C T C C A A C C T		

LeuA gene homology

	INTRON	EXON
LeuA 1 (3-1)	A T A T A C C A A T T A C C A C C T	
LeuA 1 (3-2)	A T C T A T C T T T C A C C A C C C	
LeuA 1 (3-3)	A C C A A T C C T T C A T A C C A C C A	
1	T T T T T A T A T T C C T A C C C T	
A	A T A A A T A A C C A C C C	
2	C C C C C C C C C C C C C C C C C C	

Figure 1

The homology between the three introns are shown in comparison to the LeuA gene (1) and (2) and (3) the consensus for plant introns. The plant consensus was taken from the data compiled by Altshuler et al. (1981).

different positions (51) although genes coding for higher and lower eukaryotic leucine may not be found. The predicted amino acid composition is shown in Table 1.

It is not possible to definitely assign a particular gene product to the LeuA gene but, tentatively, the predicted molecular weights of 66411 for the conserved and coded for the leucine match most closely with the LeuA protein of 66,000 (11). As the predicted leucine also matches reasonably well in terms of N-terminal sequence and amino acid sequence to the leucine described by Casey et al. (13) in a different position, the LeuA gene product probably belongs to a major, widespread, leucine sub-family.

LeuA gene homology

Comparison of the gene sequences with the sequences of several leucine genes (1, 2, 3) and the presence of three intervening sequences, two within the sequence encoding the leucine subunit and one within the sequence encoding the conserved. All three introns obey the GT-AG boundary

LeuA gene

1	T C T A T C A A T T A
2	T C T A T A A A A A

LeuA gene

1	C T C C C A A T T C T
2	C C C C C A A T T C T

LeuA gene

1	C A A G C T G C A G A A T G T C
2	T A (1-5) T T C A (2-4) C C

Figure 1

The canonical sequences of the predicted leucine 1, the sequence from the LeuA gene (2), the consensus sequence for plant systems (3), the consensus sequence for animal systems (4).

rule of Brothman et al. (12), although (13) have compared a number of sequences for the intron boundaries in higher plant genes. The leucine intron boundaries are compared (43, 44) with the plant consensus sequences and with that for animal genes (51). The leucine sequences are more in agreement with the plant consensus sequences in that the acceptor site is preceded by a run of A's rather than C's. The intervening sequences are also typical of plants (31) in being short (of leucine, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412, 413, 414, 415, 416, 417, 418, 419, 420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 485, 486, 487, 488, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 543, 544, 545, 546, 547, 548, 549, 550, 551, 552, 553, 554, 555, 556, 557, 558, 559, 560, 561, 562, 563, 564, 565, 566, 567, 568, 569, 570, 571, 572, 573, 574, 575, 576, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 597, 598, 599, 600, 601, 602, 603, 604, 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 615, 616, 617, 618, 619, 620, 621, 622, 623, 624, 625, 626, 627, 628, 629, 630, 631, 632, 633, 634, 635, 636, 637, 638, 639, 640, 641, 642, 643, 644, 645, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 658, 659, 660, 661, 662, 663, 664, 665, 666, 667, 668, 669, 670, 671, 672, 673, 674, 675, 676, 677, 678, 679, 680, 681, 682, 683, 684, 685, 686, 687, 688, 689, 690, 691, 692, 693, 694, 695, 696, 697, 698, 699, 700, 701, 702, 703, 704, 705, 706, 707, 708, 709, 710, 711, 712, 713, 714, 715, 716, 717, 718, 719, 720, 721, 722, 723, 724, 725, 726, 727, 728, 729, 730, 731, 732, 733, 734, 735, 736, 737, 738, 739, 740, 741, 742, 743, 744, 745, 746, 747, 748, 749, 750, 751, 752, 753, 754, 755, 756, 757, 758, 759, 760, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771, 772, 773, 774, 775, 776, 777, 778, 779, 780, 781, 782, 783, 784, 785, 786, 787, 788, 789, 790, 791, 792, 793, 794, 795, 796, 797, 798, 799, 800, 801, 802, 803, 804, 805, 806, 807, 808, 809, 810, 811, 812, 813, 814, 815, 816, 817, 818, 819, 820, 821, 822, 823, 824, 825, 826, 827, 828, 829, 830, 831, 832, 833, 834, 835, 836, 837, 838, 839, 840, 841, 842, 843, 844, 845, 846, 847, 848, 849, 850, 851, 852, 853, 854, 855, 856, 857, 858, 859, 860, 861, 862, 863, 864, 865, 866, 867, 868, 869, 870, 871, 872, 873, 874, 875, 876, 877, 878, 879, 880, 881, 882, 883, 884, 885, 886, 887, 888, 889, 890, 891, 892, 893, 894, 895, 896, 897, 898, 899, 900, 901, 902, 903, 904, 905, 906, 907, 908, 909, 910, 911, 912, 913, 914, 915, 916, 917, 918, 919, 920, 921, 922, 923, 924, 925, 926, 927, 928, 929, 930, 931, 932, 933, 934, 935, 936, 937, 938, 939, 940, 941, 942, 943, 944, 945, 946, 947, 948, 949, 950, 951, 952, 953, 954, 955, 956, 957, 958, 959, 960, 961, 962, 963, 964, 965, 966, 967, 968, 969, 970, 971, 972, 973, 974, 975, 976, 977, 978, 979, 980, 981, 982, 983, 984, 985, 986, 987, 988, 989, 990, 991, 992, 993, 994, 995, 996, 997, 998, 999, 1000).

At least two of these introns are conserved within the sequence of the LeuA gene of pea (data not presented). Introns have also been found in the soybean leucine (Physalis) gene (1) and these appear to be in homologous positions to the LeuA introns (10). However whilst the positions are conserved between the two species, the soybean introns are much longer. The 5' end of the gene

The perfect homology between the LeuA gene and the cDNA clone (10) suggests that the LeuA gene is transcriptionally active and has all the necessary transcriptional control sequences. A search of the LeuA sequence for the canonical subsequence observed at the 5' end of most genes from animal sources (see 13) reveals a TATA box beginning at position -40 (43, 44). This shows good homology to the consensus for the TATA box of Kuo et al. (11) and

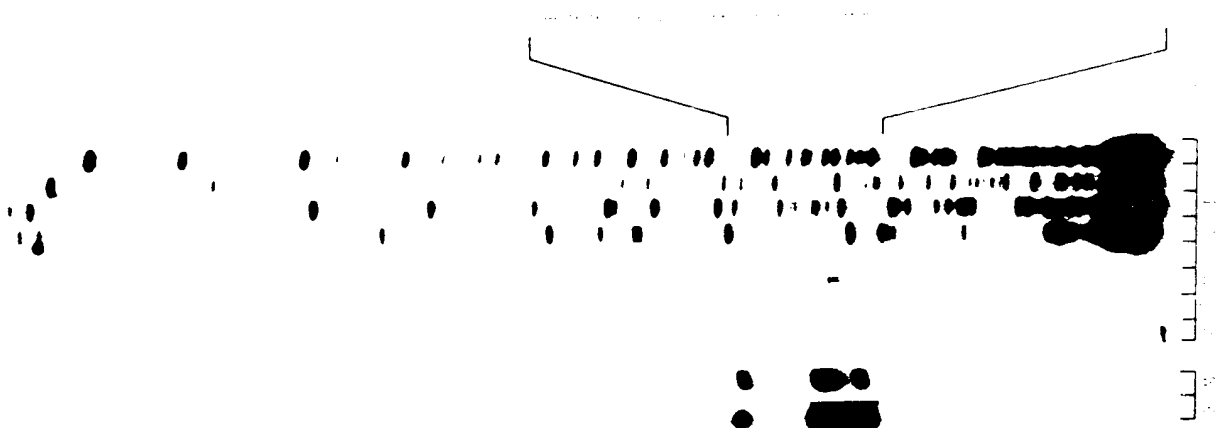
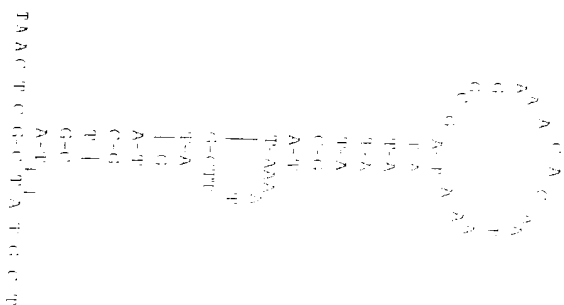


Figure 5

24 non-PCR mapping of the transfection start point. PolyA⁺ RNA was fractionated to RNA fragments and labeled at the XbaI site near to the initiation codon and extending to the polyA site directly upstream. Samples were treated with either EcoRI or XbaI. In lane 1, lane 2 (lanes w and y) or 2^{100} and z (lanes z and 2). The control lanes y and z were labeled with ten times as much c-myc as lane w and x. Lane w and x represent a 100-fold over exposure of lanes w and x respectively. A sequencing ladder prepared from the same and labeled DNA from w is run alongside (lanes 6aT and c).

(25), being preceded by the direct repeat 10 (Fig. 4). The transcriptional start point of several plant MIRNAS lie within the sequence C-IRATC/A 18-23 bp from the 'GATA box' (46) and a sequence CATT occurs 2 bp downstream from the



5. For the first time, a report of the national government, the "White Paper," has been published which lists the official responsibilities of each of the 100000 employees of the FMA, and that in parentheses, the number of persons of each ethnic group who have the status of "First Citizen" listed in Parliament. This is the first time.

'TATA box' in the *hspA* gene. The transcriptional start of the *hspA* message was located by an *in situ* nuclease digestion experiment (Fig. 5). Of the two main groups of bands which do not appear in the control, the major group bracket the *CAT* sequence while the minor group centre upon an AT pair 10 bp further downstream. Consequently, on the basis of its position, it would seem that the 'TATA box' is located at approximately -30 bp.

The 'ACAT box' is part of the promoter of the *lecA* gene. It has been observed by Messing *et al.* (15) that whilst sequences like the 'CAT box' (20) are found upstream of the 'ATA box' in animal genes and have been found in some plant genes, the homology is often poor or no 'CAT box' is apparent. A new consensus sequence known as the 'ACAT box' has therefore been proposed for plant genes (14). Examination of the *lecA* gene shows a good match for the 'CAT box' at -126 (see Fig. 1). However the sequence adjacent to this on the 5' side shows partial homology to the 'ACAT box' (Fig. 1). It is also interesting to note that a sequence showing no homology to the consensus sequence for the abbreviating enhancer core element (55) occurs 75 bp upstream of the 'CAT box' or the complementary strand (Fig. 2).

The 3' sequence of the *legA* gene shows strong homology to the 3' end of the previously published *gMA* sequences (5,15) and continues past the site of polyadenylation as found in the *Legumin gMA* sequences in *Lotus* (16,17).

31. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
32. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
33. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
34. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
35. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
36. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
37. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
38. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
39. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
40. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
41. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
42. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
43. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
44. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
45. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
46. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
47. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
48. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
49. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
50. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
51. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
52. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
53. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
54. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
55. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
56. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.
57. Leber, R., P. Berg, R. L. W. Wilson, R. P. Shochat, E. and Parkins, R. A. *Cell* 32:111-122, 1982.

A type II restriction endonuclease with an eight nucleotide specificity from *Streptomyces limbicus*

Bo-Qin Qiang* and Ira Schickel†

New England Biolabs, Inc., 32 Cove Road, Beverly, MA 01915, and *Department of Biochemistry and Molecular Biology, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, Beijing, China

Received 28 February 1984; Revised and Accepted 8 May 1984

ABSTRACT

A new site-specific endonuclease, Stl I, has been isolated from *Streptomyces limbicus*. This is the first report of a type II restriction endonuclease whose recognition specificity requires eight nucleotides. Stl I leaves the sequence, 6'-(N)N₂GAATC, symmetrically to produce a three base, 3' extension.

INTRODUCTION

The type II restriction endonucleases have become indispensable tools for molecular biology and genetic engineering. There are 91 different specificities among the known 398 restriction endonucleases (1). All require either a tetra-, penta- or hexanucleotide sequence to cleave DNA. Here we describe a new type II restriction endonuclease, Stl I, whose recognition specificity requires eight nucleotides. The purification of Stl I and determination of its recognition sequence and cleavage site are described.

MATERIALS AND METHODS

All restriction endonucleases, enzymes and DNA substrates were prepared in this laboratory. The plasmid containing the human IgM and J genes (2) was a gift of P. Leder. Alpha-³²P-deoxyadenosine triphosphate was purchased from NEN. Searches for restriction sites within known sequenced DNAs and prediction of fragment sizes were determined by the use of a computer.

Purification of Stl I. *Streptomyces limbicus* ATCC 13651 was grown at 30°C in liquid media to stationary phase. The cells were harvested and stored at -70°C. Fifty eight grams of cell paste were suspended in 5 buffer (10 mM potassium phosphate pH 7.4, 10 mM 2-mercaptoethanol, 0.1 mM EDTA) and broken by twenty 30 sec treatments with the 12.2 inch probe on a Beut Systems sonicator cell disruptor 225K. The cell debris was removed by centrifugation at 10,000 g for 20 minutes at 4°C. The supernatant, crude extract, was applied to a 2.5 x 30 cm DEAE-sepharose 6B column equilibrated with 5 buffer.

25. Freee, J.D., Price, P.W. and Metzberg, R.L. (1979) *J. Biol. Chem.* **254**, 1219-1226.
26. Hay, J., Appel, B., Schacht, J., Chan, S., Yamada H. and Söhl, O. (1982) *The Nucleic Acids Res.* **10**, 487-500.
27. Haden, B.F.H. (1982) in *The Cell (Hershey, Vol. 10, Eds. Buchb, H. and Rothman, J., pp. 319-351, Academic Press, New York.*
28. Strydom, L., Price, P.W., Rubtsov, P.M. and Kaye, A.A.A. (1979) *Arch. Acad. Biol.* **105B**, 247-261-265.
29. Strydom, L., Zahlaty, V.M., Rubtsov, P.M. and Kaye, A.A.A. (1979) *Arch. Acad. Biol.* **105B**, 247, 1275-1277.
30. Hahn, L.M. and Haden, B.F.H. (1980) *The Nucleic Acids Res.* **8**, 5993-6005.
31. Haden, B.F.H., Moss, H. and Salton, H. (1982) *The Nucleic Acids Res.* **10**, 2387-2398.

Sequence analysis of zein cDNAs obtained by an efficient mRNA cloning method

Gösta Hedrick*

Department of Biochemistry, University of California, Davis, CA 95616, USA, and

Joachim Messing

Department of Biochemistry, University of Minnesota, St. Paul, MN 55108, USA

Received 21 March 1983; Revised and Accepted 24 June 1983

ABSTRACT

A cDNA library was generated from mRNA isolated from the developing endosperm of W22 maize inbred. cDNA clones for zein, the maize storage protein family, were isolated and analyzed by DNA sequencing. The DNA sequences of four clones containing cDNA copies of mRNAs belonging to one zein subfamily were determined. The data support the following conclusions: a) genes encoding the larger of the two zein species contain eleven instead of nine repeat units within the coding sequence of the gene; b) transcription can be terminated at either of the two polyadenylation signals and c) transcription starts 31 basepairs downstream from the first T in the TATA box. To facilitate this analysis a new method for the construction of cDNA libraries was developed. The mRNA was annealed to linearized and oligo dT tailed pUC9 plasmid DNA, which then primed synthesis of the first strand of the cDNA. Oligo-dC tails were added to the cDNA-plasmid molecules, which were then centrifuged through an alkaline sucrose gradient. The gradient step removed small molecules and separated the two cDNAs which were formerly attached to the same double stranded plasmid molecule. An excess of oligo-dC tailed denatured pUC9 DNA was added and the DNA was renatured under conditions that favor the circularization of monomers by the oligo-dC and oligo-dC tails. The oligo-dC tail served as primer for the synthesis of the second strand of the cDNA. The library was screened by colony hybridization using ³²P-labelled cDNA and DNA from genomic zein clones as probes. We obtained 20,000 clones hybridizing total cDNA starting with 1 µg of plasmid DNA and 1 µg of mRNA.

INTRODUCTION

Zeins, the major storage proteins in the corn kernel, consist of a group of alcohol soluble proteins (1). They account for up to 60% of the total protein in the mature seed. Previous protein studies showed that they share a common amino acid composition (2,3) and that the amino terminus is conserved at 22 of 33 positions (4). Electrophoretically, zeins can be separated into two major (21 and 19 kd) and several minor bands (10 to 15 kd) on polyacrylamide/dodecylsulfate gels (2, 5). On two dimensional gels the proteins can be separated into at least 25 different species (6). These data suggest that the proteins are encoded by a multigene family similar to the chorion multigene family in *Drosophila* (7).

Similar to the chorion proteins, zeins are synthesized in high amounts and only in one specific organ of the organism, i.e., the endosperm of the developing kernel 18 to 52 days after pollination. During this time span the membrane bound mRNA isolated from the endosperm directs the synthesis of zeins in an *in vitro* translation system (8, 9, 10). Using cDNA clones, mRNAs were divided into 5 subfamilies by hybrid arrest translation experiments (11, 12, 13) and dot blot analysis under high stringency conditions (14, 15). Southern blot hybridization experiments with individual cDNA clones as probes showed heterogeneity within the subfamilies (16). Lowering the stringency of hybridization in a stepwise manner demonstrated that all members of the zein family show at least 60% homology (17).

To investigate the sequence divergence between and within the subfamilies we generated a cDNA library from the mRNA of the developing endosperm of W22 maize inbred. In this report we describe the generation of the cDNA library and the sequence data for four cDNA clones belonging to the same subfamily. The cDNA library was constructed by a method which combined efficiency of yield and completeness of the cDNA copies with the simplicity of previously described methods (17, 18, 19). The DNA sequence analysis of the four cDNA clones reveals that genes within one subfamily can encode both the 19 and 21 kD proteins. The length of the peptide chain depends upon the number of repetitive units which have been previously found to account for most of the structural part of zein genes (20, 21, 22, 23). Furthermore, the sequence data suggests that the start of transcription occurs 31 basepairs downstream from the first T in the TATA box and that there is heterogeneity in respect to which of the two polyadenylation signals, frequently found in plant genes (24), is utilized.

MATERIALS AND METHODS

Strains

The pUC plasmids, the M13mp phage vectors and the strains JM103 and JM83 have been described elsewhere (25, 26).

Media, Transformation, Histochemical assay, Chemicals, and Enzymes

Media and plates were as described by J. Miller (27). Transformation of competent cells was as described by Cohen *et al.* (28) except that the CaCl_2 concentration was raised to 50 mM. Cells were transfected with M13mp phage and plated in soft agar with 5x10⁸ fresh cells in the presence of 1 mM IPTG and 0.004% X-gal. Cells transfected with pUC plasmids were allowed to grow for 40 min in the absence of selection before they were streaked on YT plates

containing 100 µg/ml ampicillin, 1 mM IPTG and 0.004% X-gal. IPTG was omitted when JM83 was used.

IPTG (isopropyl-β-D-galactopyranoside) was obtained from Sigma, X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside) from Bachem, dNTPs and ddNTPs from Fl biochemicals, radiolabelled dNTPs from Amersham and RNasin from biotec. All enzymes were either purchased from Bethesda Research Laboratories or New England Biolabs except for terminal deoxynucleotidyl transferase (terminal transferase) and reverse transcriptase, which were obtained from Ratliff Biochemicals and Dr. J. Beard, respectively.

Tailing Reactions

20 µg of pUC9 DNA were digested with *Pst*I and the degree of digestion was monitored by agarose gel electrophoresis. The DNA was extracted with phenol and phenol/chloroform and ethanol precipitated twice to ensure the removal of residual phenol. The precipitate was washed with 70% ethanol and dried under vacuum for 15 min. The DNA was resuspended in 20 µl ddH₂O and divided into two aliquots. For the oligo-dT-tailing reaction 10 µl 1 M K-cacodylate pH 7.0, 23 µl H₂O, 0.5 µl 0.1 M DTT, 2 µl 1 mM *P*-dNTP, 1.5 µl terminal transferase (16U/µl) and 5 µl 10 mM CoCl_2 were added in the listed order. The components for the oligo-dC-tailing reaction were the same except that 2 µl 0.5 mM *P*-dCTP were used instead of the dTTP. The mixtures were incubated at 37°C for 30 min. The RNA was phenol extracted and ethanol precipitated three times to remove the Co^{2+} which would interfere in the reverse transcriptase reaction. The tailed plasmid DNAs were finally resuspended in 10 µl of low Tris buffer (10 mM Tris pH 7.6, 10 mM NaCl, 1 mM EDTA).

To tail the cDNA-plasmid conjugates the DNA was resuspended in 10 µl ddH₂O, 4 µl 1 M K-cacodylate, 1 µl 1 mM dGTP, 1 µl 0.05 M DTT, 2 µl 20 mM MnCl_2 (29) and 1 µl terminal transferase (16U/µl). The reaction was incubated for 15 min at 37°C. The DNA was phenol extracted, ethanol precipitated and dissolved in 50 µl low Tris buffer.

cDNA Synthesis

The cDNA synthesis reaction was done in a final volume of 15 µl and incubated 90 min at 37°C. The components were: 800 µM dATP, dCTP, dGTP, dTTP, 70 mM KCl, 50 mM Tris pH 8.2, 10 mM MgCl_2 , 2 mM DTT, 1 U/µl RNasin, 25 µg/ml actinomycin D, 40 nM oligo-dT-tailed pUC9 DNA (1 µg), 250 nM RNA, 100 U/ml reverse transcriptase. The poly A-mRNA and the oligo-dT-tailed plasmid DNA anneal under these conditions during the reaction. After the reaction the DNA was phenol extracted once and ethanol precipitated three times. To prevent the precipitation of unincorporated nucleotides the ethanol precipi-

tates were warmed to room temperature before centrifugation.

Alkaline Sucrose Gradient Centrifugation

A 5 ml linear sucrose gradient was used (5-20% sucrose w/v in 0.2 M NaOH, 0.8 M NaCl, 1 mM EDTA with a 0.5 ml 60% sucrose cushion). The sample was diluted with 50 μ l of the 5% sucrose solution and layered on the gradient. Centrifugation was carried out in a SW 50.1 rotor at 36k rpm for 17 hr at 4°C. The gradient was collected from the bottom in 0.3 ml fractions. The profile of the gradient was established based on the ⁶⁰Co gamma radiation of the fractions.

Reannealing and Ultraclearization of the cDNA-Plasmids

The amount of plasmid in the pooled fractions was calculated based on the relative amount of radioactivity and oligo-dT-tailed pUC9 DNA was added in 5 to 10 fold excess. The solution was then dialyzed against low Tris buffer in the cold to remove the NaOH and NaCl. The DNA was concentrated in the presence of 25 μ g/ml carrier RNA and resuspended in 50 μ l low Tris buffer. Concentrated NaCl, Tris pH 8., formamide and ddH₂O were added to give final concentrations of 1-5 μ g/ml plasmid DNA, 32% (v/v) formamide, 50 mM NaCl, 10 mM Tris (30, 31). The annealing mix was incubated for 24 hr at 37°C, dialyzed against 100 mM NaCl, 10 mM Tris pH 8.0, 1 mM EDTA in the cold and concentrated by ethanol precipitation.

ELISA Reaction

The annealed DNA was taken up in 50 μ l of cold 50 mM NaCl, 20 mM Tris pH 7.6, 10 mM MgCl₂, 1 mM DTT, 100 μ M dATP, dGTP, dTTP. 3 units of DNA polymerase I-large fragment were added and the mixture was incubated for 60 min at 15°C and 60 min at room temperature. The DNA was phenol extracted, ethanol precipitated and resuspended in 50 μ l low Tris buffer.

Isolation of Maize Endosperm mRNA

Maize kernels were harvested 22 days after pollination and mRNA was extracted as described by Burr *et al.* (9). The mRNA was size fractionated by centrifugation through a 5-20% sucrose DMSO gradient and purified subsequently by passage through an oligo-dT cellulose column. The RNA was assayed for biological function in a cell free translation system as described by Park *et al.* (12).

DNA Sequencing

The cDNA inserts were sequenced using the chain termination method (32) and the M13 subcloning procedure (33) except that a synthetic universal primer was used. A detailed protocol for the subcloning strategy and sequencing procedure is presented elsewhere (34). Storage and processing of

the sequencing data was conducted with the Apple II Microcomputer (35, 36).

Electrophoresis

Sequencing gels were essentially prepared as described elsewhere (37), except that for long runs 5% gels (acrylamide/bisacrylamide 40:1) were used. Agarose gel electrophoresis and visualization of nucleic acids were carried out as described elsewhere (33). Denaturing agarose gels contained 5 mM OH⁻ and were prepared in 50 mM boric acid, 5 mM Na₂B₄O₇, 10 mM Na₂SO₄, 1 mM EDTA Na₂, pH 8.2.

RESULTS AND DISCUSSION

Construction of the cDNA Library

The major steps of the procedure are outlined in Fig. 1. The mRNA is annealed to oligo-dT-tails which had been added at the PstI site of pUC9 plasmid DNA. The oligo-dT-tails prime the cDNA synthesis along the mRNA. The plasmid-cDNA conjugates are in turn extended with oligo-dG-tails, denatured and sized by centrifugation through an alkaline sucrose gradient. Molecules of the appropriate length are renatured under dilute conditions in the presence of an excess of oligo-dC-tailed, single-stranded pUC9 DNA. The low DNA concentration favors the formation of monomers circularized by hybridization between the dC- and dG-tails. The second strand of the cDNA is primed by the oligo-dC-tail and synthesized by DNA polymerase I-large fragment.

Step 1: Preparation of Primer Vector and Second Strand Vector

pUC9 plasmid DNA was prepared as previously described (25) and cleaved with restriction endonuclease PstI. The tailing reactions were carried out as described in Materials and Methods. The reactions were monitored in three ways. The total number of nucleotides incorporated per molecule of plasmid was calculated from the proportion of TCA precipitable radioactivity. To determine the length of the individual tails an aliquot of the tailing reaction was digested with EcoRI. This releases one tail per plasmid molecule which can be visualized by electrophoresis through a denaturing polyacrylamide gel followed by autoradiography. The resulting picture showed a ladder (data not shown). Based on a sequencing reaction of a known template which was run in parallel the steps of the ladder corresponded to chains of 65 to 85 nucleotides for the dT-tailing reaction and 45 to 65 for the dC-tailing reaction. The actual tails are shorter by the distance between the EcoRI and PstI site which is 26 bp. The integrity of the plasmid DNA was tested by digesting another aliquot with HaeIII. The fragments were

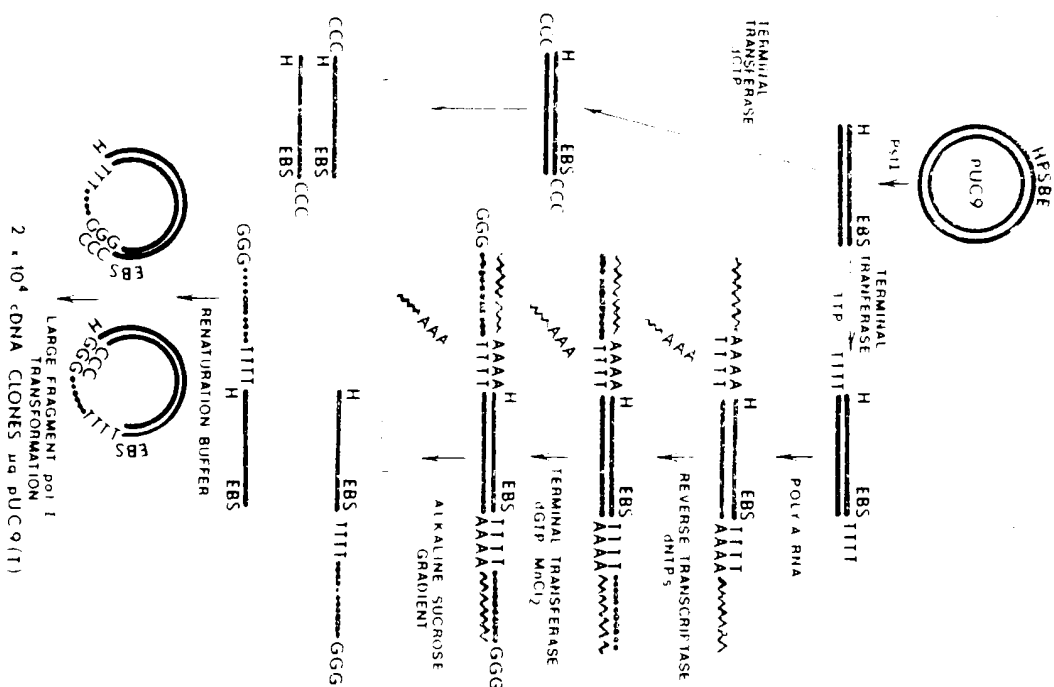


Figure 1: Flowsheet for the cDNA cloning procedure. Explanations are given in the text. E, B, S, and H stand for the EcoRI, BamHI, SalI and HindIII sites in the multicloning site of pUC9.

separated into four bands on a 1.5% agarose gel. The smallest two bands had shifted up in comparison to a HaeIII digest of PstI cleaved pUC9 DNA (Fig. 2). The autoradiograph of this gel showed that greater than 90% migrated with the lower two bands. Higher labelling of the other bands would have indicated that the plasmid DNA had been nicked before or during the tailing reaction to

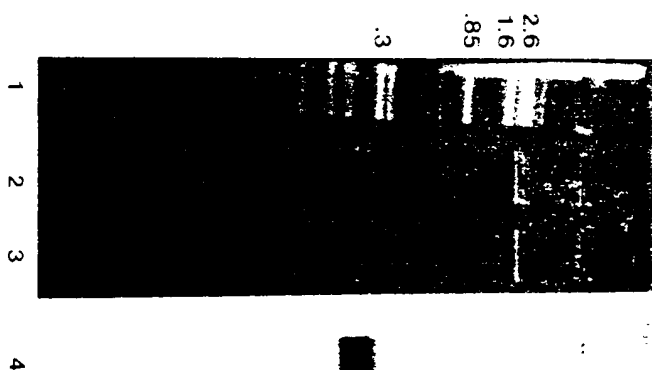


Figure 2: Analysis of oligo-dT-tailed pUC9 DNA. Homopolymer tails of deoxythymidilate were added to PstI cleaved pUC9 DNA. An aliquot of this DNA was digested with HaeIII and the resulting fragments were separated on a 1.5% agarose gel (lane 3). The autoradiograph (lane 4) of the gel reveals that >90% of the radiolabel resides in the smallest two fragments. These two fragments show a decreased mobility when compared to non-tailed pUC9 DNA cleaved with PstI and HaeIII (lane 2). Lane 1 shows M13mp2 Rf DNA digested with HaeIII, the sizes of some fragments are given.

a degree that would interfere with the following steps in the procedure.

Step 2: cDNA Synthesis

mRNA from developing seeds of corn inbred W22 was isolated as described in Materials and Methods. The mRNA was converted into cDNA using the oligo-dT-tailed plasmid DNA as primer (see Materials and Methods). A three fold molar excess of mRNA had previously been determined to saturate the system. Under these conditions about 60% of the tails primed cDNA synthesis as judged by comparing equimolar amounts of oligo-dT-tailed plasmid and plasmid-cDNA conjugates on a denaturing agarose gel (data not shown). The length of the cDNA transcripts can be estimated from the same gel.

Step 3: Addition of oligo-dG tails

The cDNA-plasmid molecules were tailed in the presence of dGTP and $MnCl_2$.

This combination results in optimal (though not equally efficient) tailing of any kind of end (29). There is no easy way of monitoring this reaction. As both the molar concentration (see below) and the kind of ends in the reaction are unknown the incorporation of labelled dGTP does not allow a calculation of the average tail length. It was, therefore, necessary to perform this reaction under conditions where both dITP and terminal transferase were present in excess. Under these conditions the length of the tails depended on the incubation time and the conditions given in Materials and Methods resulted in the addition of 10 to 25 residues of dGMP. This had been established in a control experiment previously and was verified by sequencing data.

Step 4: Sizing and Strand Separation of cDNA-Plasmids

After the de-tailing reaction the molecules were fractionated on an alkaline sucrose gradient (see Materials and Methods). The vector and cDNA-vector molecules banded in a well defined peak separated by four fractions from a gradually rising slope of small molecules (Fig. 3). The gradient served four purposes:

- Enrichment of molecules longer than the vector itself. This will increase the proportion of clones that contain cDNA inserts over those that just contain "tails".
- Elimination of the short DNA molecules. These small molecules seem to be generated during the cDNA synthesis step and may be primed by RIA fragments or by the poly-A tail of the mRNA hybridized to the

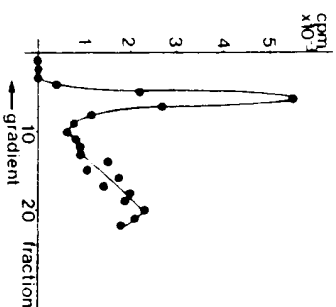


Figure 3: Profile of cDNA-plasmid conjugates on an alkaline sucrose gradient.

Oligo-dT-tailed pUC9 DNA was used to prime cDNA synthesis with maize endosperm mRNA serving as template. The cDNA-plasmid molecules were extended with terminal transferase in the presence of dGTP and subjected to alkaline sucrose gradient centrifugation as described in Materials and Methods. Sedimentation is from right to left.

primer vector. As these molecules also carry oligo-dG tails they would interfere in the following step, especially as their molar concentration is difficult to determine.

- Removal of the RIA by base hydrolysis.
- Separation of the two strands of plasmid DNA and, thus, the two cDNAs which were formerly attached to the same double-stranded vector molecule. This is necessary in order to obtain plasmid molecules with a single cDNA insert.

Step 5: Reconstitution and Circularization of the Plasmid

To reconstitute a double-stranded circular plasmid molecule with only one cDNA insert, 3 µg of oligo-dC tailed pUC9 DNA were added to the pooled fractions 4 to 6. This represented a seven to ten fold excess of dC-tailed plasmid over de-tailed cDNA-plasmid. Base and salt were removed and the DNA was reannealed in the presence of formamide. The conditions given in Materials and Methods result in greater than 90% reannealing of the DNA as judged from an agarose gel on which the position of the reannealed material was compared to native and denatured pUC9 DNA (data not shown). The formamide was gradually removed by dialysis to allow circularization of molecules with dC- and dG-tails. As the DNA concentration was low, circularization was more likely to occur than the formation of concatemers. Finally, the DNA was concentrated by ethanol precipitation and the second strand of the cDNA was filled in by large fragment of DNA polymerase I.

Step 6: Screening of the cDNA Library

Transformation of 2.5% of the final product of the cloning procedure gave about 500 white and 100 blue colonies. The blue colonies most likely arose from uncleaved vector DNA present in the oligo-dC tailed preparation. The white colonies were screened for the presence of cDNA inserts by colony hybridization using ³²P-labelled, randomly primed cDNA or nick translated DNA from two genomic Zein clones representing two subfamilies (38). About 80% of the clones hybridized to the cDNA probe, and about 20% hybridized to the two combined zein probes (data not shown). Colonies that only hybridized weakly to the Zein probes were later shown to contain cDNAs belonging to a different subfamily. Extrapolating the number of clones that hybridized to the various probes, we estimate that 30 to 40% of our clones contained Zein cDNAs. This proportion is surprisingly low as Zein proteins were by far the most prevalent products when the mRNA used in this experiment was translated in vitro (8, 9, 10, 11). To determine the size of the inserts plasmid DNA from 60 white colonies was isolated and sized by agarose gel electrophoresis.

Twelve had inserts of about 1200 bp, 7 carried short inserts of 100 to 200 bp and the remaining 41 had inserts between 400 and 900 bp (data not shown).

COMMENTS

The method presented here describes a scheme that allows the synthesis and cloning of cDNA copies of mRNA in only 6 steps. The three enzymes that are used in the procedure are all commercially available. Except for the oligo-dG-tailing reaction of the cDNA-plasmid conjugates every reaction in the scheme can be easily monitored and thus controlled. The length of the cDNA inserts primarily depends on the length of the mRNA. In the experiment reported here 400 to 1200 nucleotide long mRNAs were used and both the length of the first strand cDNA copies, as well as the length of the inserts in the final clones, reflected this size distribution. In a pilot experiment using 9S RNA from rabbit reticulocytes a large proportion of the cDNA-plasmid conjugates were of uniform length and indicated that first strand cDNA transcripts of 550 nucleotides were the major product. On the bases of length and restriction pattern, about 50% of the cDNA clones isolated in this experiment had inserts that were full length cDNA copies of globin mRNAs (data not shown). First strand cDNA copies of mouse mammary tumor virus RNA prepared under these conditions again showed the same size distribution as the RNA (1000 to 9000 nucleotides) (data not shown). The only step in the procedure that might reduce the length of the cDNA and still result in its being cloned is the oligo-dG-extension of the cDNA-plasmid conjugates. This can be prevented by using terminal transferase that is free of nuclease activity.

Even though any plasmid that contains a cloning site with 3' protruding ends could be used with this protocol, the pUC plasmids are particularly suitable. Since the PstI site is flanked by several other restriction sites the cDNA insert can be released by one (pUC7) or two (pUC8, 9, 12, 13) restriction enzymes. Furthermore, by choosing the right pUC plasmid the procedure can be adapted to experimental goals that go beyond the simple cloning of cDNA. In the scheme presented here cDNA inserts in both orientations are obtained. By removing one of the oligo-dT-tails with the appropriate restriction enzyme before the first strand cDNA synthesis all the resulting clones should contain cDNA copies in either the sense or nonsense orientation in respect to the lac promoter. This allows the introduction of defined deletions from either the 5' or 3' end of the insert in subsequent experiments. Insertion of the cDNA in the sense orientation should also result in the expression of fusion proteins between the aminoterminal end of

beta-galactosidase and cDNA encoded peptides in some of the clones. This would allow one to screen the library by immunological assays and make it possible to isolate genes for which only the protein has been isolated so far (40).

Since the M13mp vectors contain the same array of restriction sites as the pUC plasmids, sequences cloned in pUC plasmids can be shuttled between these two vector systems and the cloned DNA can be easily isolated as single stranded DNA. This is of use for the preparation of strand specific hybridization probes (41), site specific mutagenesis with oligonucleotides (42) and for DNA sequencing by the chain termination method (32). If the cDNA is cloned directly into the M13mp vectors for sequencing, then the universal primer directs DNA synthesis through a run of dA or dG (33). Depending on the length of these homopolymers, the Sanger sequencing reaction (32) is inhibited (not shown). Most sequencing studies require at least one subcloning step and, thus, we do not feel that it is necessary to use the M13mp vectors as the primary cloning vehicle in the cDNA-plasmid procedure. There are several important reasons for using the pUC plasmids instead of the M13mp vectors. First, pUC is a smaller vector which allows a better separation of the cDNA-plasmid conjugates from the vector molecules. Second, pUC plasmids are not restricted to a male host. Thus, bacterial strains that exhibit high transformation efficiencies can be used (43). Although strains with the M15 deletion either in the F'-traJ36 or in the *E. coli* chromosome, are host strains (44) which allow pUC plasmids with inserts to be monitored by a simple color reaction (25), this test is usually superfluous when used with recombinant DNA libraries. If the library has to be amplified before screening, M13 phage containing inserts have a severe growth disadvantage when compared to "empty" phage, which is in contrast to lambda and plasmid libraries (45).

Sequence Analysis of cDNA Clones

Four Zein cDNA clones that hybridized labelled Z4 DNA (38) were isolated. Based on the restriction maps of these clones, Sau3A, AluI, RsaI and EcoRI* were used to fragment the cDNA inserts. The fragments were subcloned into M13mp vectors both in a shotgun fashion and by forced cloning (46, 26). The M13mp subclones were sequenced by the chain termination method (32).

Figure 4 shows a comparison of the four cDNA sequences Z67 and Z619, Z631 and Z6124 to the sequence of the genomic clone Z4 (21). Also included in the comparison is the sequence of cDNA clone A30 (20, 47), which is the

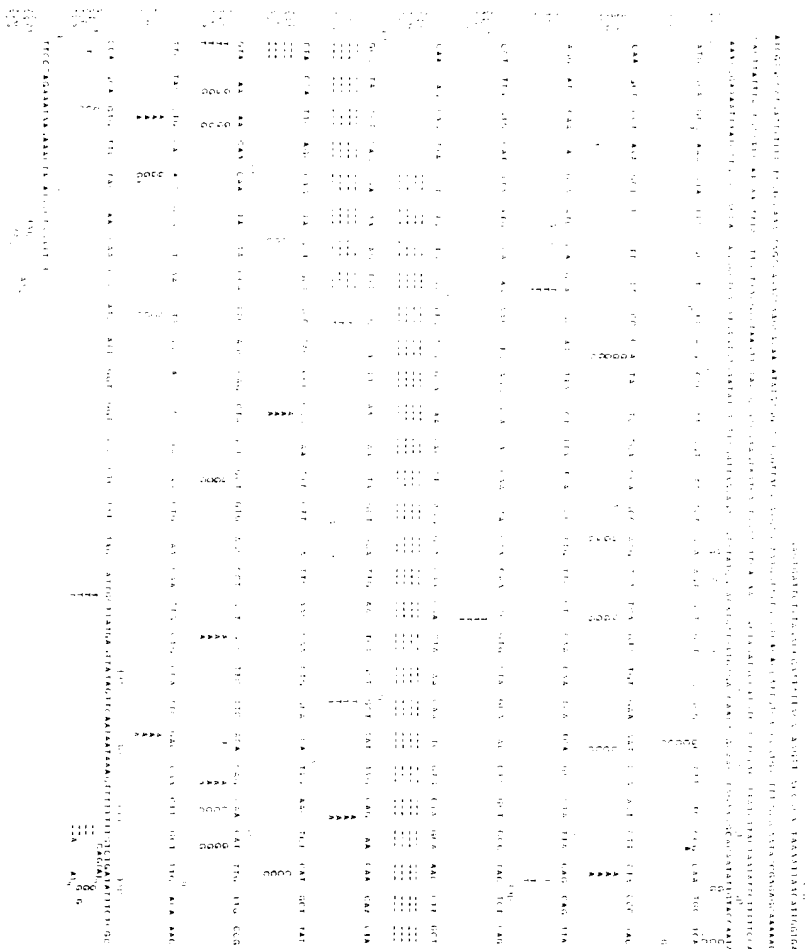


Figure 4: Comparison of the nucleotide sequences of cDNA clones belonging to the Z4 family of zeln genes. The sequences are numbered starting with the first transcribed nucleotide. The translated part of the sequence is written in triplet form. Dashes indicate nucleotides not present in cDNA clones Z6124, A30, Z631 and Z619. The complete sequence of genomic clone Z4 is presented, but only the variable nucleotides for the cDNAs are given at the site of their occurrence. The sequences of Z4, A30 and Z631 were previously published (21).

copy of a mRNA isolated from a different inbred line (Illinois High Protein). Z67 and Z6124 are probably full length cDNA clones as they have a common start which is located 31 bp downstream from a consensus TATA sequence in the genomic clone Z4 (Z4). Two other cDNA clones (Z614 and Z615; Heidecker and Messling, in preparation) which belong to a different subfamily start at this position, whereas all other cDNAs sequenced vary at their 5' ends. Both Z6124 and Z67 start with a Guanylyl residue which does not match the genomic sequence. This G-residue could be due to contaminating nucleotides either in

the de-tailing reaction of the cDNA-plasmid or in the de-tailing reaction of the plasmid. Alternatively, it could be the beginning of a loop back at the end of the cDNA. The latter explanation seems more likely as none of the tails show any evidence of contaminating nucleotides. Moreover, we observed the same GATC palindrome at the start of the other two full length cDNA sequences (Z614 and Z615), which were mentioned above.

The comparison shows that overall the sequences are indeed very highly conserved the lowest degree of homology between any two clones being 95%. However, even this closely related subfamily can again be divided into two subgroups. cDNA clones Z619, Z631, Z6124 and A30 share most of the 40 differences in comparison to the genomic clone Z4. Any pairwise comparison shows at least 99% homology. cDNA clone Z67 and Z4 share an internal duplication of 96 basepairs and an extra codon in comparison to the other four clones. These two sequences differ by only 10 single nucleotide exchanges. The extensive collinearity between Z4 and Z67 identifies Z4 as an active gene and supports our previous data (21) showing that our genomic clone does not contain any intervening sequences but represents a member of the larger size class of zeln genes. Because of the lack of matching cDNA and genomic clones a more indirect approach, including electron microscopy and S1 mapping, has been used in other studies to document the absence of intervening sequences in zeln genes (22, 48).

Z67 and Z619 share a remarkable feature. Both clones originated from mRNAs that terminated after the first of the two polyadenylation signals frequently found in plant genes (24). The distance between the signal and the site of polyadenylation is about the same for all five cDNA clones. Thus, plant genes may contain either one or two polyadenylation signals and in at least some cases where two signals are present, either signal can be recognized. This situation has been found for only a few animal genes (49, 50, 51). It remains to be elucidated whether this feature is used to regulate or control gene expression at the mRNA transcription and/or processing level, or whether it is just a random event without any particular consequences.

The proteins encoded by the 6 closely related genes described in this report represent both major zeln size classes. The molecular weights predicted from the nucleotide sequence all are higher by about 4 kd than those determined by SDS/polyacrylamide electrophoresis. The latter values most likely are underestimates due to the hydrophobicity of the proteins. The molecular weights of the proteins deduced from the sequences of Z4 and

Z67 are 29.2 kd for the precursor and 27 kd for the mature protein. The sequences of Z6124 and A50, and probably of Z631 and Z619, can likewise be translated into proteins of 25.4 kd and 23.3 kd for the two forms. These data do not support a report by Marks and Larkins (15) who found that each subfamily, as defined by cross-hybridization under stringent conditions, directed the synthesis of proteins of only size class. These conflicting results may be caused by differences in the inbred line, mRNA isolation, and/or peculiarities of in vitro translation experiments. Park *et al.* (12) reported translation arrest experiments which suggested that clone A50 hybridized with mRNA that coded for both the higher and the lower mol. wt. species of proteins in most of the inbred lines used in these studies. Taking into account the variables described above, the hybridization data of the latter report agree very well with our sequencing data. Although 2 out of 6 cloned sequences of this subfamily code for the larger protein, a quantitative representation of both size classes cannot be calculated, because these numbers are too small to allow a valid statistical analysis. On the other hand, sequencing 5 clones from the same subfamily in the same inbred line did not result in any identical sequences, which gives us an indication that this subfamily alone must be far more complex than 5 members.

ACKNOWLEDGEMENTS

We would like to thank Dave Pratt, John Ingraham, Bob Cardiff, Irwin Rubenstein and Thomas Fanning for their help and discussion throughout the work and Ida Fierro and Kris Kohn for their aid in preparing this manuscript. This research was supported by the Department of Energy, DE-AC02-81ER 10901, the National Institutes of Health, St32 GM07467-05, and the Minnesota Exp. Station, MN-15-030.

*Permanent address: Department of Pathology, School of Medicine, University of California, Davis, CA 95616, USA

REFERENCES

1. Wall, J. S. and Paulis, J. W., in Pomeranz, Y (ed.), *Adv. Cereal Science and Technol.*, Vol. 11. Am. Assoc. Cereal Chem. St. Paul, Minn, pp. 135-219, 1975.
2. Lee, K. H., Jones, R. A., Dalby, A. and Tsai, C. (1976) *Biochem. Genet.* 14, 641-650.
3. Glanazza, E., Vliglenghi, V., Rhiigetti, P. G., Salamini, F. and Soave, C. (1977) *Phytochemistry* 16, 315-317.
4. Bletz, J. A., Paulis, J. W. and Wall, J. S. (1979) *Cereal Chem.* 56, 327-332.
5. Glanazza, E., Rhiigetti, P. G., Ploil, F., Galante, E. and Soave, C. (1976) *Maydica* 21, 1-17.
6. Hagen, G. and Rubenstein, I. (1980) *Plant Sci. Lett.* 19, 217-223.
7. Etkubsh, T. H., Jones, C. W. and Kafatos, F. C. In Brown, D. D. (ed.), *194-1113 A Symposium on Molecular and Cellular Biology*, Vol. XXIII. Academic Press, New York, pp. 135-153, 1981.
8. Larkins, B. A., Jones, R. A. and Tsai, C. Y. (1976) *Biochemistry* 15, 5506-5511.
9. Burr, B., Burr, F. A., Rubenstein, I. and Simon, M. N. (1978) *Plant Sci. Lett.* 13, 365-375.
11. Wlenand, U., Bruschke, C. and Felix, G. (1979) *Nucl. Acids Res.* 6, 2707-2715.
12. Park, W. D., Lewis, E. D. and Rubenstein, I. (1980) *Plant Physiol.* 65, 98-106.
13. Violiti, A., Abildsten, D., Pogna, N., Sala, E. and Pirodda, V. (1982) *EMBO J.* 1, 53-58.
14. Burr, B., Burr, F. A., St. John, T. P., Thomas, M. and Davis, R. W. (1982) *J. Mol. Biol.* 154, 33-49.
15. Marks, M. D. and Larkins, B. J. (1982) *J. Biol. Chem.* 257, 9976-9983.
16. Hagen, G. and Rubenstein, I. (1981) *Gene* 13, 239-249.
17. Maniatis, T., Kee, S. G., Efstratiadis, K. and Kafatos, F. C. (1976) *Cell* 8, 163-182.
18. de Marfynoff, G., Pays, E. and Vassart, G. (1980) *Biochem. biophys. Res. Comm.* 93, 645-653.
19. Okayama, H. and Berg, P. (1982) *Mol. Cell. Biol.* 2, 161-170.
20. Geraghty, D., Pelfer, M. A., Rubenstein, I. and Messing, J. (1981) *J. Nucl. Acids Res.* 9, 5163-5174.
21. Hu, N.-T., Pelfer, M. A., Heldecker, G., Messing, J. and Rubenstein, I. (1982) *EMBO J.* 1, 1337-1342.
22. Pedersen, K., Devereux, J., Wilson, D. R., Sheldon, E. and Larkins, B. A. (1982) *Cell* 29, 1015-1026.
23. Argos, P., Pedersen, K., Marks, M. D. and Larkins, B. (1982) *J. Biol. Chem.* 257, 9984-9990.
24. Messing, J., Geraghty, D., Heldecker, G., Hu, N.-T., Kridl, J. and Rubenstein, I., in Hollaender, A., Kosuge, T. and Meredith, (eds.), *Genetic Engineering of Plants*. Plenum Press, New York, pp 211-228, 1983.
25. Vieira, J. and Messing, J. (1982) *Gene* 19, 259-268.
26. Messing, J. and Vlierira, J. (1982) *Gene* 19, 269-276.
27. Miller, J. Experiments in molecular genetics. Cold Spring Harbor Laboratory, Cold Spring harbor, NY, 1972.
28. Cohen, S. N., Chang, A. C. Y. and Hsu, L. (1972) *Proc. Natl. Acad. Sci. USA* 69, 2110-2114.
29. Deng, G. and Wu, R. (1982) *Nucl. Acids Res.* 9, 4172-4188.
30. Fanning, T. G., Schreier, P. F. and Davies, R. W. (1976) *Eur. J. Biochem.* 62, 173-170.
31. Dugalizcyk, A., Boyer, H. W. and Goodman, H. (1975) *J. Mol. Biol.* 96, 171-184.
32. Sanger, F., Nicklen, S. and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463-4567.
33. Heldecker, G., Messing, J. and Gronenborn, B. (1980) *Gene* 10, 69-73.
34. Messing, J., in Wu, R. (ed.), *Methods in Enzymology*, Vol. 101. Academic Press, New York, pp. 29-78, 1983.
35. Larson, R. and Messing, J. (1982) *Nucl. Acids Res.* 10, 39-49.
36. Larson, R. and Messing, J. (1983) *DNA* 2, 31-35.
37. Sanger, F. and Coulson, A. R. (1978) *FEBS Lett.* 87, 107-110.

38. Lewis, E. R., Hagen, G., Mullins, J. L., Mascia, P., Park, W. D., Banton, W. D. and Rubenstein, I. (1981) *Gene* 14, 205-215.
39. Holmes, D. S. and Quigley, M. (1981) *Anal. Biochem.* 114, 193-197.
40. Heffman, D. M., Feramisco, J. R., Fiddes, J. C., Thomas, G. P. and Hughes, S. H. (1983) *Proc. Natl. Acad. Sci. USA* 80, 31-35.
41. Hu, H.-T. and Messing, J. (1982) *Gene* 17, 271-277.
42. Zoller, M. J. and Smith, M., in Wu, R. (ed.), *Methods in Enzymology* Vol. 101, Academic Press, New York, in press.
43. Morfiscio, D. A., in Wu, R. (ed.), *Methods in Enzymology*, Vol. 68, Academic Press, New York, pp. 326-331, 1979.
44. Messing, J. *Recombinant DNA Technical Bulletin*, NIH Publication No. 79-99, 2, No. 2 (1979) 43-48.
45. Messing, J., in Setlow, J. and Hollaender, A. *Genetic Engineering*, Vol. 4, Plenum Publishing Company, New York, pp. 19-35, 1982.
46. Messing, J., Iida, R. and Seeburg, P. H. (1981) *Nucl. Acids Res.* 9, 309-321.
47. Geraghty, D., Messing, J. and Rubenstein, I. (1982) *EMBO J.* 1, 1329-1335.
48. Wienand, U., Langridge, P. and Felix, G. (1981) *Mol. Gen. Genet.* 182, 440-444.
49. Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R. and Hood, L. (1980) *Cell* 20, 313-319.
50. Setzer, D. R., McGeehan, M., Nurnberg, J. H. and Schimke, R. T. (1980) *Cell* 22, 361-370.
51. Tosi, M., Young, R. A., Hagenbuchle, O. and Schibler, V. (1981) *Nucl. Acids Res.* 9, 2313-2323.
52. Grunstein, M. and Hogness, D. S. (1975) *Proc. Natl. Acad. Sci. USA* 72, 3961-3965.

Sequence homologies in the protamine gene family of rainbow trout

J.M. Allen, D.McKenzie, H. Z. Zhou, T.C. Soares and G.H. Dixon

Department of Medical Biochemistry, Faculty of Medicine, Health Science Centre, University of Calgary, Calgary, Alberta, T2N 4N1, Canada

Received 1 March 1983; Revised and Accepted 24 June 1983

ABSTRACT

We have sequenced five different rainbow trout protamine genes plus their flanking regions. The genes are not clustered and do not contain intervening sequences. There is an extremely high degree of sequence conservation in the coding and 3' untranslated regions of the genes. Bowstern sequences exhibit little homology though conserved regions are found 250 base pairs 3' to the genes. There are four regions upstream of the gene that are highly conserved in the six clones, including the canonical Goldberg-Hogness box which is 35 base pairs 5' to the coding region. A second homology region is found 90 bases upstream. Although in the same approximate location as the CAAT box found upstream of other genes, it does not contain the canonical CAAT sequence. Further upstream of the protamine genes at -115 there is an A-T rich sequence while a 25 base pair conserved sequence is located 120 bases upstream. In addition we report the presence of a potential 7 MBZ region of predominantly A-C repeats approximately one kilobase downstream of one of the genes.

INTRODUCTION

Evolutionarily conserved RNA sequences have been strongly implicated in the regulation of eukaryotic gene expression. Comparison of the 5' flanking regions of various genes has resulted in the discovery of consensus sequences (reviewed by 1), such as the Goldberg-Hogness box (25 base pairs upstream of the RNA initiation site) and the CAAT box (approximately 80 base pairs 5' to the gene). Evidence of the specific function of these regions in the transcriptional process has been somewhat equivocal. In vitro studies indicate only the Goldberg-Hogness box is necessary for faithful transcription (2,3). In vivo experiments, however, demonstrate that both the Goldberg-Hogness and the CAAT box are required for efficient transcription (4,5,6). The Goldberg-Hogness box is necessary for the specificity of initiation, while the CAAT region affects transcriptional efficiency. The sea urchin histone H2A gene seems to be an exception, as the deletion of the CAAT region does not affect the rate of transcription (7). Sequences far from the mRNA initiation site have also

- bley, A., and Zachary, W., *Biochim. J.*, **115**, 599-604, 1975.
30. Linnarsson, D., *Plant Physiol.*, **65**, 125-130, 1980.
31. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
32. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
33. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
34. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
35. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
36. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
37. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
38. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
39. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
40. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
41. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
42. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
43. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
44. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
45. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
46. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
47. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
48. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
49. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
50. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
51. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
52. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
53. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
54. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
55. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
56. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
57. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
58. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
59. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
60. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
61. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
62. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
63. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
64. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
65. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
66. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
67. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
68. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
69. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
70. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
71. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
72. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
73. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
74. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
75. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
76. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
77. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
78. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
79. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
80. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
81. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
82. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
83. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
84. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
85. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
86. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
87. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
88. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
89. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
90. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
91. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
92. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
93. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
94. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
95. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
96. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
97. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
98. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
99. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.
100. Linnarsson, D., and Zachary, W., *Plant Physiol.*, **65**, 125-130, 1980.

Structural sequences are conserved in the genes coding for the α - and β -subunits of the soybean 7S seed storage protein

Mary A. Schuler¹, Beth L. Adair², Joseph C. Pollack¹, Gregory J. Havel¹, and Roger N. Beachy^{1*}

¹Plant Biology Program, Department of Biology, Washington University, St. Louis, MO 63130, and
²Biochemistry Department, University of Missouri, Columbia, MO 65212, USA

Received 25 June 1982; Revised 28 September 1982; Accepted 11 October 1982

ABSTRACT

Cloned mRNAs encoding four different proteins have been isolated from recombinant cDNA libraries constructed with cDNA from seed mRNAs. Two cloned mRNAs code for the α and β -subunits of the 7S seed storage protein (conglycinin). The other cloned mRNAs code for proteins which are synthesized *in vitro* as 69,000 d., 69,000 d., or 69,000 d. polypeptides. Hybrid selection experiments indicate that, under low stringency hybridization conditions, all four cDNAs hybridize with mRNAs for the α and β -subunits and the 69,000 d., 69,000 d., and 69,000 d. *in vitro* translation products. Within three of the mRNAs, there is a conserved sequence of 155 nucleotides which is responsible for this hybridization. The conserved nucleotides in the α and β -subunit cDNAs and the 69,000 d. polypeptide cDNAs span both coding and noncoding sequences. The differences in the coding nucleotides outside the conserved region are extensive. This suggests that selective pressure to maintain the 155 conserved nucleotides has been influenced by the structure of the seed mRNA. RNA blot hybridizations demonstrate that mRNA encoding the other major subunit (β) of the 7S seed storage protein also shares sequence homology with the conserved 155 nucleotide sequence of the α and β -subunit mRNAs, but not with other coding sequences.

INTRODUCTION

Literature on the expression of the genes for the legume seed storage proteins has been accumulating rapidly. The studies deal with a variety of legumes, including glycyne max (soybean), *Phaseolus vulgaris* (French garden bean) and *Pisum sativum* (garden pea), and include characterization of storage protein complexes by sucrose gradient fractionation (1,2), the storage protein subunits by peptide mapping (3,4,5) and characterization of the mRNAs for the storage proteins by *in vitro* translation assays (3,4,6,7,8,9). From this work, two major classes of storage proteins referred to as the legumin (11S sedimentation coefficient) and the vicillin (7S sedimentation coefficient) (2) have been identified in most legumes. Both the 7S and 11S classes of storage proteins contain a number of closely related major subunits (3,5,10). The similarities in the subunit organization and the amino acid compositions of the various legumin and vicillin

holoproteins were used by Iwabe et al. (2) to suggest that the peptide sequences we selected for the construction, stability and/or utilization of the storage proteins are conserved within the 11S and 7S classes of proteins. Data to support this suggestion has not been presented.

The major subunits of the 7S storage protein in soybean *max* are designated as α' (83,000 d.), α (76,000 d.) and β (53,000 d.) (11,12). The amino acid compositions (13) and the proteolytic cleavage fragments (3) of the α , α' and β subunits suggest that the individual 7S subunits do contain regions of homology. In order to delineate the regions of conservation in the storage protein subunits, we have characterized cloned soybean seed cDNAs that have sequence complementarity with the mRNAs of several different 7S subunits. In the accompanying paper (14), we demonstrate that the α and α' subunits encoded by two of these cDNAs are nearly identical from the midpoint of their polypeptides to their carboxyl-termini. In this paper, hybridization of segments from the cloned cDNAs encoding the α and α' subunits to seed mRNAs reveals that the β -subunit mRNAs contain only those sequences which correspond to the carboxyl terminal coding sequences of the α and α' subunit mRNAs. The implications of this amino acid homology for protein structure are discussed.

We have also characterized two other cloned cDNAs which share a restricted region of homology with the α , α' and β -subunit mRNAs. These cDNAs encode members of an abundant class of seed mRNAs whose initial translation products are 68,000 d., 60,000 d. and 53,000 d. The region of nucleotide conservation in the mRNAs for the 7S subunits and one of the mRNAs in this second class (p68-mRNA) encompasses the same region of homology that exists between α , α' and β subunit mRNAs. DNA sequence analysis indicates that the region of nucleotide conservation is translated into amino acids present in the α and α' -subunits but not in the p68-polypeptide. Because regions of amino acid homology do not exist in the two classes of seed proteins, it appears that these nucleotides have been conserved because they play an internal role in the expression, structure or stability of the seed mRNAs.

MATERIALS AND METHODS

The first cDNA library, containing Hind III-linked double-stranded cDNAs, was constructed and screened as outlined in Beachy et al., (12). Construction of the second cDNA library containing poly(dA) tailed double stranded DNAs is described in the accompanying paper (14). The procedures

for DNA blot hybridizations and restriction site mapping by partial endonuclease digestion of end-labeled DNA fragments are detailed in Scholer et al. (14). The *in vitro* translation of soybean seed RNAs in wheat germ extracts were done according to Beachy et al. (12). The procedures for the hybrid selection of specific mRNA sequences from total soybean poly (A)⁺ RNA are described in Vercan et al. (15).

The molecular weights of soybean poly (A)⁺ RNAs complementary to the cDNA clones were determined after transfer of RNAs from 1% agarose gels containing 10 mM methylmercuric hydroxide, to activated diazobenzyloxymethyl paper (16), and hybridization with 3²P-labeled probes for 24 hr at 47°C in 50% formamide, 0.75 M NaCl, 0.075 M Na citrate (5x SSC), 0.04% Ficoll, 0.04% polyvinylpyrrolidone, 0.05% bovine serum albumin, 0.3% sodium dodecyl sulfate (SDS), 10 μ g/ml sonicated calf thymus DNA, 40 mM sodium phosphate (pH 6.5), 0.1 μ g/ml poly A. Blots were washed four times in 1x SSC containing 0.1% SDS at 37°C and exposed to Kodak XAR-5 film for 2 days with intensifying screens.

End labeled probes for the RNA blots were prepared by labeling Hind III restriction sites in α and α' 236 and α and p68.232 (Fig. 2) with reverse transcriptase and α -³²P dNTPs as described in Maxam and Gilbert (17). The resulting end-labeled DNAs were cleaved with Hae III and sized on 4% acrylamide gels. The Hind III-Hae III restriction fragments were eluted from the gel and used directly for hybridization without concentration. The purity of the labeled probes was assessed by hybridizing them to blots containing Hind III-Hae III restriction fragments of α and α' 236 and α and p68.232.

The method of Maxam and Gilbert (17) as modified by Smith and Calvo (18) was used for sequencing DNA. All sequences were carried through at least two sets of sequencing reactions. Dot matrix analyses of nucleotide homologies were carried out with computer programs similar to those shown in Konkel et al. (19).

RESULTS

Identification of cDNA Clones containing Sequences for the α and α' -subunits mRNAs. Two libraries of cloned soybean cDNAs were screened for sequences homologous to the 7S subunit mRNAs present in soybean seed embryos. Construction of the first library, by the ligation of Hind III-linked double stranded cDNAs into the Hind III site of pTR262 (20) and the subcloning into pBR322 (21) has been described in detail by Beachy et al. (12). Clones from the first cDNA library were screened directly for

sequences complementary to the α and β' subunit mRNAs by hybrid selection (15). The α and β' subunit mRNAs were isolated from this first selection and shown in this paper. The second selection prevented the amount of non-cloned mRNAs sequences the mRNAs having the closest sequence homology with the cloned cDNA.

The 750 base pair cloned cDNA, α cDNA 230 (Fig. 1) hybrid selects mRNAs for the α subunit in *in vitro* and α' subunit in *in vivo* subunits and for the β subunit in *in vitro* and β' subunit in *in vivo* subunits. *In vitro* and *in vivo* selection with α cDNA 230 have been shown to hybrid select mRNAs that encode a presumed α subunit (17). The mRNA selection products of the α cDNA 230 hybridization presented here and in nearly equal (17) indicate that the α cDNA 230 mRNAs complex dissociates at temperatures at least 5°C higher than those required to dissociate the α' mRNAs, β mRNAs and β' mRNAs hybrids. These results indicate that the cloned cDNA 230 encodes an α subunit mRNA and contains sequences present in the mRNAs for the α and β' subunits as well as the β and β' in *in vitro* and *in vivo* subunits. In addition, selection of the cloned cDNA 230 (Fig. 1) gave by *in vitro* and *in vivo* hybrid selection the α and β' subunit mRNAs and the β and β' subunit mRNAs (Fig. 1B), but has the greatest homology with the β subunit peptide mRNA.

Additional cloned cDNAs were obtained from a second cDNA library constructed by the insertion of poly(AA)-poly(AA) tailed double-stranded soybean cDNAs into the Hind III site of pBR322 (18). This library was screened for sequences complementary to the α and β' subunit cDNAs by hybridization of 32P-labeled cDNA 230 (Fig. 1) to RNA blots containing restriction fragments from each cloned cDNA. Two cDNA clones selected by this method are α cDNA 32 and β' cDNA 326 (other cDNA clones selected on the basis of their hybridization to the α cDNA 230 probe are described in the accompanying paper (19)).

As shown in Fig. 1B, the 1300 bp long cloned cDNA α cDNA 32 hybrid selects mRNAs for the α and β' subunits and the β and β' subunit peptides in a manner similar to the α cDNA 230. While the mRNAs for the β and β' subunit peptides elute from the 1300 bp long cDNA 32, the α subunit mRNAs do not elute from cDNA 32 at 70°C. In contrast to the elution profiles in the α cDNA 230 selection experiments, the majority of the α subunit mRNA hybrids dissociate between 70°C and 85°C. α cDNA 32 appears to be a cDNA encoding the α subunit that has a high degree of sequence complementarity with both the α and β' subunit mRNAs. DNA sequence analysis presented later in this paper supports this

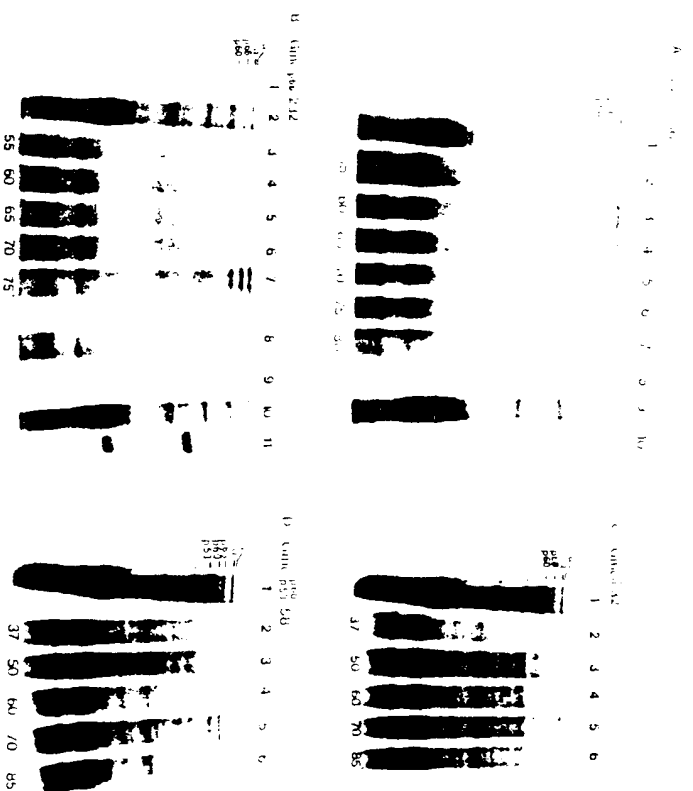


Figure 1. Hybrid selection experiments with cloned soybean seed cDNAs. (A) Purified plasmid cDNAs were denatured, bound to nitrocellulose filters and hybridized with total poly(AA)⁺ RNA isolated from mature soybean seeds. The bound mRNAs were eluted at increasing temperatures and translated in an *in vitro* wheat germ translation system utilizing ³⁵S-leucine. The resulting polypeptides were electrophoresed on to acrylamide gels and analyzed by fluorography. (A) Hybrid selections with cDNA 230 pHA.

(1) and (9) total *in vitro* translation products with seed poly(AA)⁺ RNA; (2) translation products of mRNA eluted at 55°C; (3) eluted at 60°C; (4) eluted at 65°C; (5) eluted at 70°C; (6) eluted at 75°C; (7) eluted at 80°C; (8) *in vivo* labeled mature 7S soybean proteins; (10) endogenous translation products. The relevant polypeptides are designated on the left. (B) Hybrid selections with cDNA 326 pHA. (1) *in vivo* labeled mature 7S proteins; (2) and (10) total *in vitro* translation products with seed poly(AA)⁺ RNA; (3) translation products of mRNA eluted at 55°C; (4) eluted at 60°C; (5) eluted at 65°C; (6) eluted at 70°C; (7) total *in vitro* translation products; (8) hybrid selections with cDNA 32 pHA. (1) total *in vitro* translation products with seed poly(AA)⁺ RNA; (2) translation products of mRNA eluted at 37°C; (3) eluted at 50°C; (4) eluted at 60°C; (5) eluted at 70°C; (6) eluted at 85°C.

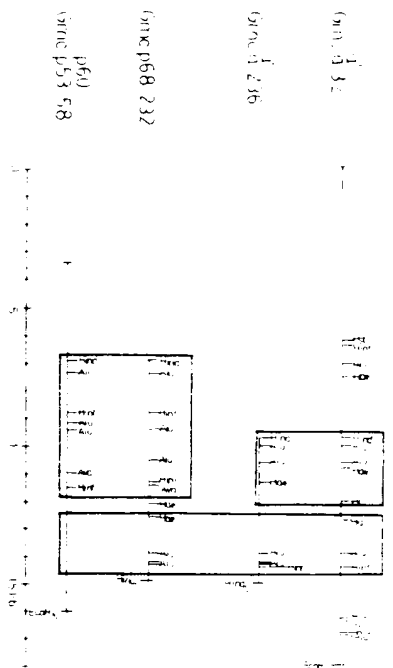


Figure 2. Restriction endonuclease cleavage sites in the cloned soybean seed cDNAs. The restriction endonuclease sites were determined by partial digestion of end-labeled pNA fragments (14). The direction of transcription for the cloned cDNAs, *Gene d'236*, was determined from the pNA sequence (Fig. 3). The *Gene d'32*, *Gene p68-232* and *Gene p53-58* restriction site maps have been aligned with the region of *Gene d'236* with which the 7' consubstituted.

The large open box delineates the region conserved in the four cloned cDNAs; the small open boxes characterize the regions of homology shared by the *Gene d'236* and *Gene d'32* cDNAs or by the *Gene p68-232* and *Gene p53-58* cDNAs. The restriction enzyme abbreviations used: Bgl (Bgl II), Hae (Hae III), Hinf (Hinf I), Acl (Acl II), Pst (Pst I), Alu (Alu I) and Hind (Hind III).

Conclusion

The result of hybrid-selection experiments with *Gene p53-58* (Fig. 10), indicate that the 1450 bp long insert of this clone hybridizes with the α and α' -subunit mRNAs and the *p68* and *p60*-polypeptide mRNAs. In addition, *Gene p53-58* binds mRNA encoding an *in vitro* translation product designated as *p53* (53,000 d, *in vitro*). The mRNAs encoding the *p53* and *p60*-polypeptides elute between 60°C and 70°C while other mRNAs elute at temperatures less than 30°C. Thus, *Gene p53-58* contains sequences complementary to the mRNAs for the *p60* and *p53*-polypeptides and shares limited homology with the mRNAs for the α and α' -subunits and the *p68*-polypeptide.

In summary, the hybrid selection experiments indicate that these four cloned cDNAs have different sequence complementarities with the mRNAs for two and possibly three of the 7S storage protein subunits and for the *p68*, *p60* and *p53* *in vitro* translation products.

Restriction Analysis of the Cloned cDNAs Encoding the α and α' -Subunits and the *p68*, *p60* and *p53* Polypeptides. The initial step in determining the regions of homology in the four cloned cDNAs was the construction of the fine structure restriction endonuclease maps shown in Fig. 2. The sites for the frequently cutting restriction enzymes in these maps were defined by

partial endonucleolytic digestion of end-labeled DNA fragments, as described in Schuler et al. (14). Comparison of the four restriction maps indicates that some restriction site homology exists between *Gene d'236* and *Gene d'32*, the cloned cDNAs which strongly select α and α' -subunit mRNAs. *Gene p53-58* site similarities also exist between the *Gene p68-232* and the *Gene p53-58* cloned cDNAs. The only set of the restriction sites in *Gene p68-232* that matches those in *Gene d'236* is the closely spaced triplet of Alu I restriction sites at the 3' end of the *Gene d'236* pNA.

The regions of nucleotide homology between the cloned cDNAs (Fig. 2) were determined by blot hybridization of restriction fragments from *Gene d'32*, *Gene p68-232* and *Gene p53-58* DNA with the end-labeled 200 bp and 350 bp Hind III-the III subfragments of *Gene d'236* and the 530 bp and 300 bp Hind III-the III subfragments of *Gene p68-232*. The 300 bp Hind III-the III subfragment of *Gene d'236* hybridized only with the *Gene d'32* restriction fragments (data not shown). The 530 bp Hind III-the III subfragment of *Gene p68-232* hybridized only with restriction fragments of *Gene p53-58*, even at low hybridization stringencies. In contrast, cleavage fragments from all four cloned cDNAs hybridized with the 350 bp Hind III-the III subfragment of *Gene d'236*. We conclude that part of the sequence within the 350 bp Hind III-the III subfragment of *Gene d'236* is shared by all of these cloned DNAs and that sequences upstream from the conserved region are shared exclusively by the *Gene d'236* and *Gene d'32* DNAs, the α/α' unique sequences, and by *Gene p68-232* and *Gene p53-58* DNAs, the *p68*, *p60* and *p53* unique sequences. Sequences shared by all of these cloned cDNAs are referred to as the "conserved 7S protein mRNA sequences".

Sequence Analysis of the Cloned cDNAs. Sequence analysis of each cloned cDNA in the region complementary to the 350 bp Hind III-the III subfragment of *Gene d'236* is shown in Fig. 3. The *Gene d'236* and *Gene p68-232* cDNAs do not contain the full length 3' noncoding sequences or the poly(A) tracts of the mature mRNA. Results described in the accompanying paper (14) demonstrates that the 3' border of the *Gene d'236* pNA lies 31 nucleotides upstream from the termination codon of the α -subunit mRNA. The 3' border of *Gene p68-232* DNAs lies in the 3' noncoding region of the *p68*-mRNA.

The proteins encoded by *Gene d'236* and *Gene d'32* (Fig. 3) are nearly identical and have amino acid compositions similar to the α and α' -subunit proteins (Table 1; 13). The sequence analyses and partial amino acid analyses of the cloned cDNAs that are discussed in the accompanying paper (14) indicate that *Gene d'236* represents an α -subunit mRNA and *Gene d'32* represents an α' -subunit

8252

Figure 3. The nucleotide sequences for the cDNAs encoding the α and α' -subunits and the p68 and p69 polypeptides. The nucleotide sequence of the cloned cDNAs, *tmca* 32, *tmca* 230, *tmca* p68-232 and *tmca* p69-238 are shown on lines 2, 3, 6 and 7. The nucleotides included in the brackets at the center of the *tmca* 32 are derived from the closely related cDNA, *tmca* 10 (14). Undetermined nucleotides are designated by X. The amino acids encoded by *tmca* 32 are given on line 1; the amino acids encoded by *tmca* 230 which differ from these are shown on line 4. The amino acids encoded by *tmca* p68-232 are presented on line 5; the amino acids encoded by *tmca* p69-238 which differ from these are shown on line 8. Solid vertical lines (|) between two sequences mark the nucleotides that are identical. The conserved nucleotides are enclosed in a box.

Table 1. Comparison of the amino acids coded for by the cloned cDNAs, Gm α 23b, Gm α 32, Gm p68-232 and Gm p53-58. The amino acids coded by the four cloned seed cDNAs are shown in columns 2-5. The number of residues calculated in mole percents are shown in parentheses. The mole percent amino acids in the full length mature α (7b), α (32), α (53,58) and β (53,58) d.) subunits (13) are listed in columns 6-8.

Amino Acid	Residues in Cloned cDNA				Mole Percent in Mature		
	Gm α 23b	Gm α 32	Gm p68-232	Gm p53-58	Subunits		
Aspartic acid	8 (4.9)	12 (4.0)	4 (2.4)	4 (2.4)	12.21	11.75	13.35
Glutamic acid	13 (7.9)	18 (6.0)	11 (7.2)	12 (7.2)	31.00	26.01	23.01
Alanine	15 (9.1)	29 (9.7)	5 (3.5)	6 (3.6)			
Valine	17 (10.3)	27 (9.0)	10 (7.0)	11 (6.6)			
Leucine	11 (6.7)	21 (7.0)	7 (4.9)	9 (5.4)	4.41	5.24	4.49
Arginine	5 (3.0)	16 (5.3)	8 (5.6)	6 (3.6)	6.50	5.47	5.65
Glutamine	13 (7.9)	17 (5.7)	11 (7.7)	11 (6.6)	1.11	2.57	1.45
Proline	13 (7.9)	19 (6.3)	16 (11.3)	18 (10.4)	4.35	4.25	3.19
Isoleucine	19 (12.2)	26 (8.7)	12 (8.2)	19 (11.5)	2.06	3.10	3.15
Threonine	12 (7.3)	22 (7.3)	6 (4.2)	7 (4.2)	4.25	4.23	4.51
Phenylalanine	9 (5.5)	11 (3.7)	5 (3.5)	6 (3.6)	7.02	6.65	5.88
Tryptophan	3 (1.8)	13 (4.3)	7 (4.9)	8 (4.8)	4.09	4.08	4.43
Methionine	0	2 (0.7)	1 (0.7)	2 (1.2)			
Cysteine	1 (0.6)	2 (0.7)	0	0	0.34	0.53	0.35
Serine	3 (1.8)	16 (5.3)	4 (2.8)	11 (6.6)	6.45	6.05	6.47
Threonine	10 (6.1)	30 (10.0)	10 (7.0)	11 (6.6)	4.09	4.33	5.21
Glutamine	1 (0.6)	6 (2.0)	5 (3.5)	7 (4.2)	2.21	2.35	3.16
Proline	0	0	1 (0.7)	1 (0.6)	0	0	0
Tyrosine	3 (1.8)	7 (2.3)	2 (1.4)	9 (5.4)	2.16	2.38	2.39

and α -subunit cDNAs, Gm α 32 and Gm α 23b, differ from each other at 12 nucleotide positions within the 155 bp conserved sequence. In spite of the extensive nucleotide conservation, the cloned cDNAs do not encode the same amino acids in the α -subunits and the p68/p66/p53 polypeptides. The conserved regions of the α and α -subunit mRNAs are translated into amino acids which lie near the carboxyl-terminus; the conserved region of the p68 mRNA spans the noncoding portion of the mRNA.

The nucleotide homologies of the two cloned cDNAs Gm α 32 and Gm p68-232 have been evaluated by dot matrix analysis (Fig. 4) to determine if other nucleotide sequences are shared by these DNAs. From this analysis, it is clear that the sequences of Gm p68-232 and Gm α 32 DNA are homologous only in the 155 bp conserved region. The coding sequences upstream from the conserved region of the α -subunit and p68-polypeptide mRNAs can not be aligned even if insertions and deletions are artificially introduced into the sequences. The homologies between the Gm p68-232 and Gm p53-58 DNAs and between Gm α 23b and Gm α 32 are so extensive that presentation of their dot matrices would be redundant. It should be noted that in these comparisons there was no evidence of reiterated sequence elements within the α -subunit DNA or the p68-polypeptide DNA.

Hybridization of the Conserved and Nonconserved Sequences of the Cloned

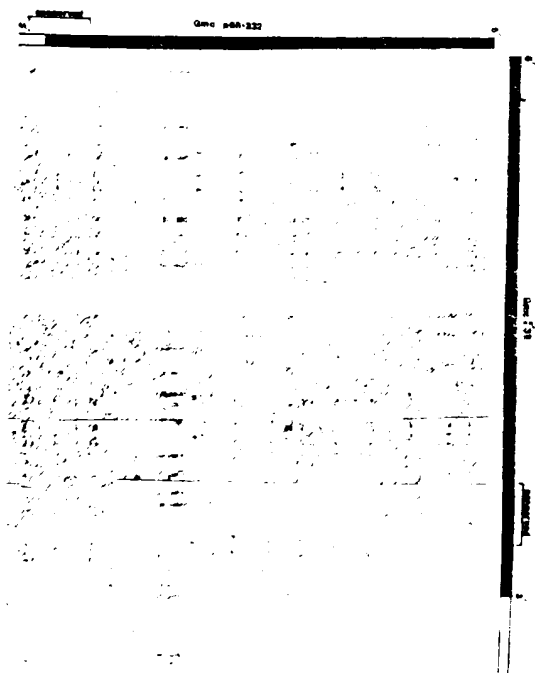
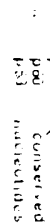


Figure 4. Dot matrix comparison of the cloned soybean cDNAs for the α -subunit and the p68-polypeptide. The nucleotide sequences of the cloned cDNAs Gm α 32 and Gm p68-232, presented in Fig. 3, have been compared by dot matrix analysis (19,22). In the sequence comparisons plotted here, each dot corresponds to the first nucleotide in a series of six for which at least four nucleotides match. The diagonal of the plot represents a base-for-base alignment of the two sequences; diagonal sets of dots which lie on either side of the center diagonal correspond with homologous sets of nucleotides that are aligned with one another by nucleotide insertions or deletions. The coding (black boxes) and noncoding (open boxes) nucleotides of the two cloned DNAs are designated at the top and left of the figure; the 155 nt. long conserved sequences are bracketed.

cDNAs to soybean seed RNAs. The molecular weights of mRNA species complementary to the conserved and unique regions of the α -subunit cDNA, Gm α 23b, and the p68-polypeptide cDNA, Gm p68-232, were determined by the RNA blot method of Alwine et al. (16). The soybean seed mRNAs shown in Fig. 5A were hybridized with the α -subunit mRNA specific sequences contained in the 200 bp Hind III-Hae III subfragment of Gm α 23b (Fig. 2). Fig. 5B demonstrates that this fragment contains sequences complementary to the 2500 nucleotide long α and α -subunit mRNAs.

The p68 polypeptide mRNA specific sequences present in the 530 bp Hind III-Hae III subfragment of Gm p68-232 hybridize with mRNAs 2500 nucleotides in length (Fig. 5C). These mRNAs are sufficient in size to code for the 68,000 d., 60,000 d. and 53,000 d. *in vitro* translation products. It has already been demonstrated that the sequences represented in the 530 bp



sequence and the amount could be derived from the RNA blot hybridization of soybean coryledon and axis poly(A)⁺ RNAs to the conserved region sequences. These tissue specific mRNAs display striking differences in their hybridization patterns with the conserved sequence probe (Fig. 5b). The 1700 nt. mRNA species which hybridizes with conserved sequences in the 390 bp Hind III-Bcl II fragment of cDNA 236 is present only in coryledon mRNA and absent from axes mRNA. Thus, in conjunction with our previous report that cDNA 236 DNA hybrid selects 3'-subunit mRNA (12), strongly suggests that the 1700 nt. mRNA codes for the 3'-subunit and contains the conserved nucleotides. We have not determined whether the region of homology between the α and α' -subunit mRNAs and the 3'-subunit mRNA is limited to the 155 bp conserved region shared by cDNA 236 and cDNA p08-252. We know only that the boundaries of the homology in the

DISCUSSION

We have characterized four cloned cDNAs which are complementary to several soybean seed mRNAs. Hybrid-selection experiments identify the soybean seed mRNAs with the closest sequence homology to each cloned DNA. These results together with sequence analysis of the four cDNAs indicate that two cDNAs code for the α and β subunits of the 20S protein.

storage complex. The other two cDNAs code for proteins which have primary translation products of 68,000 d., 60,000 d. or 53,000 d. The derived amino acid sequences show that the members of this latter group of proteins are related.

We have also shown that all four cloned cDNAs hybridize to different

extents with mRNAs for the α and α' -subunits, as well as those for the p68, p60 and p53 polypeptides. DNA sequence comparisons revealed that three of the four cloned mRNAs share a highly conserved region of 155 nucleotides. Aside from these nucleotides, the mRNAs coding for the α and α' -subunits have few nucleotides in common with those coding for the p68, p60 and p53 polypeptides. The sequence conservation within this 155 nt. conserved region is high (32-99%), and contrasts strongly with the sequence variation in the remainder of the mRNAs for the α and α' -subunits and the p68 polypeptides.

Surprisingly, the conserved region is positioned differently relative to the formation codons for each class of mRNAs. As a consequence of this, the conserved nucleotides are translated into amino acids situated 31-83 residues from the carboxyl-terminus of the α and α' -subunit proteins, and are not translated into amino acids in the 68,000 d. polypeptide. Thus, the nucleotide conservation does not seem to be the result of conservation in the carboxyl-terminal amino acids in these seed proteins. It appears more likely that the selective pressure to maintain the conserved set of nucleotides has been influenced by the structure of the mRNAs. A primary sequence or secondary structure in these mRNAs may be conserved so that expression of these genes can be regulated in the steps between transcription and production of mature polypeptides, e.g., by altering the stability or translational efficiency of the mRNA. The absence of the conserved nucleotides from the 600 p53 mRNA is puzzling and further characterization of other p68 and p60-polypeptide mRNAs will indicate whether any other mRNAs in this class lack the conserved nucleotides.

The biological significance of the 68,000 d., 60,000 d. and 53,000 d. polypeptides is not known, nor have the size of the mature proteins derived from these primary translation products been determined. The mRNAs encoding these polypeptides are present at the same developmental stages (1-0) as the mRNAs for the α , α' and β -subunits (R.N. Beachy, unpublished results). In an earlier study, several seed proteins ranging in size from 58,000 d. to 70,000 d. were shown to be present in the early and middle stages of seed maturation (1-0) (26), and were shown by pulse chase studies to be synthesized in maturing seed embryos (R.N. Beachy, unpublished results). Furthermore, the 53,000 d., 60,000 d. and 68,000 d. translation products and the 58,000 to 70,000 d. seed proteins are precipitated with antisera directed against the seed proteins sedimenting at 7S in sucrose density gradients (B.F. Iadla and R.N. Beachy, unpublished results). Thus, it

appears that the 58,000 d. to 70,000 d. group of proteins may be the mature forms of the *in vitro* translation products designated as p53, p60 and p68. Because protein processing steps for these polypeptides have not been studied, we do not know the relationship between the polypeptides produced *in vitro* and those produced *in vivo*.

The immunoprecipitation experiments cited above suggest either that antigenic similarities exist between the α , α' and β subunits and the p68, p60 and p53 polypeptides or that the mature proteins derived from the p68, p60 and p53 polypeptides form a holoprotein that sediments with a density of 7S. Because of the absence of amino acid homology between the α and α' -subunits and the p68, p60, p53 polypeptides in the portions of the proteins presented in this paper, it is unlikely that the two classes of polypeptides share common antigenic determinants. Monoclonal antibodies directed against the individual seed proteins are needed to determine if the mature products derived from the p53, p60 and p68 polypeptides associate into a 7S holoprotein which is different from the 7S conglycinin storage protein or if they provide the nucleating structures for the formation of the 7S conglycinin holoprotein.

The conserved nucleotides which are shared by the α and α' -subunit mRNAs and the p68, p60 and p53 polypeptide mRNAs are also present in mRNA for the β -subunit of the 7S seed storage protein. Whether the region of nucleotide conservation in the α , α' and β -subunit mRNAs is limited to the 155 nucleotides shared by the p68-polypeptide and α -subunit mRNAs or whether it includes as many as 350 nucleotides is unknown. Because the β -subunit mRNA does not share extensive nucleotide homology with the unique regions of either gene α' 236 or gene p68*232, the β -subunit must be a encoded by a separate gene(s) which shares little amino acid homology with the α and α' -subunit genes.

If the same reading frame is used for the translation of the conserved nucleotides in the α , α' and β -subunit mRNAs, then similar amino acids exist in all three subunits. The sequences of the α and α' -subunits have been subjected to secondary structure analysis using the rules developed by Chou and Fasman (27,28) and Garnier et al. (29) (M.A. Schuler, unpublished results). The results of these analyses suggest that the amino acids encoded by the conserved regions in the α and α' -subunits participate in the formation of three antiparallel β -pleated sheets. Other experiments will help to determine where the β -pleated sheet regions encoded by the conserved nucleotides are positioned within the 7S holoprotein and whether the amino

Closely related families of genes code for the α - and α' -subunits of the soybean 7S storage protein complexMary A. Schindler, Eric S. Schmidt and Roger N. Beachy¹

Plant Biology Program, Department of Biology, Washington University, St. Louis, MO 63130, USA

Received 25 June 1982; Revised 28 September 1982; Accepted 11 October 1982

ABSTRACT

Thin-clone cloned cDNA encoding the α and α' -subunits of the 7S seed storage protein in the soybean, *Glycine max*, have been isolated from a recombinant cDNA library constructed with mRNA from maturing seeds. In addition, a gene encoding an α' -subunit has been isolated from a recombinant Charon 4A phage library containing genomic *Glycine max* DNA. The cloned DNAs have been divided, on the basis of their endonuclease sites, into two main classes of sequences which differ in approximately 6% of their nucleotides. Whereas the proteins encoded within each DNA class are nearly identical, the proteins encoded by the two different classes of soybean DNAs are distinct and correspond to α and α' -subunits. Thus, the α and α' -subunits are coded for by two closely related multigene families. The amino acid differences in the portions of the α and α' -subunits presented in this paper occur primarily near the carboxyl-terminus. The 3' noncoding nucleotides of the cloned α and α' -subunit DNAs are more highly conserved than are the coding nucleotides. This conservation suggests that the 3' untranslated sequences of the α and α' -subunit mRNAs are functional in the expression of the α and α' -subunit proteins or in the stabilization of the 7S subunit mRNAs.

INTRODUCTION

The abundant synthesis of the soybean (*Glycine max*) seed storage proteins in a period confined to the developmental stages of cell enlargement and seed maturation (1) provides an excellent opportunity for studying the regulation of gene expression in a higher plant. Conglycinin, which constitutes one of the two major seed storage protein complexes in the soybean, sediments in sucrose gradients as a 7S protein complex of three subunits. The trimeric components of the 7S protein peak are formed from various combinations of the three major subunits [β (83,000 d.), α (76,000 d.), β (53,000 d.) of conglycinin (2,3)]. The individual α' and α subunits show a high degree of similarity in their amino acid content (4,5) and in their proteolytic cleavage fragments (6). In addition to the peptide similarities between the subunits, each type of subunit in the 7S protein complex of the mature seed can be resolved into multiple components

(four α' , three to four β , two β' species) on two dimensional isoelectric focusing-SDS gels (R.E. Jantz and R.R. Beachy, unpublished results). Thus, the proteins which constitute the 7S seed storage protein complex exhibit several levels of peptide homology.

The amino acid conservation among the subunits and the multiplicity of subunit isotypes has suggested that each major subunit is encoded by a small gene family that is closely related to the gene families for the other subunits of the 7S soybean seed storage protein. Before the existence of a distinct small gene family for each 7S subunit could be demonstrated, the genes for the three types of 7S subunits were shown to be related at the nucleotide level in mRNA hybrid selection experiments with a cloned seed cDNA that hybridized to α' and β' subunit mRNAs (3). Further experiments discussed in Schubert *et al.* (7) have indicated that sequences within a 350 bp region of the cloned β' subunit cDNA are conserved in the genes for each of the major 7S subunit proteins. These findings are consistent with the estimate of 5-20 copies/genome of the 7S subunit genes reported by Goldberger *et al.* (8) in liquid hybridization experiments using sequence probes common to the α , α' and β subunit mRNAs. Because of the repetitive nature of the gene families for the individual 7S subunits and the homology between the gene families, studies characterizing the supragenital organization of the individual gene families must precede experiments examining the expression of the 7S storage protein genes and the genetic engineering of these plant genes.

In this paper, we identify cloned cDNAs containing sequences complementary to the mRNAs encoding the α and α' subunits of the 7S storage protein and elucidate the overall organization of the closely related α and α' subunit gene families. The nucleotide homologies existing between members of these gene families indicate that they are highly conserved throughout the 3' terminal half of their coding sequences and their 3' noncoding sequences. The nucleotide differences in the coding regions of these mRNAs indicate that the α and α' subunit mRNAs encode distinct proteins which differ primarily in the amino acids near the carboxyl-termini. Variations in the amino acids within each family of cDNAs potentially can be correlated with the numerous isotypes of the α and α' subunits expressed in the seed. In addition, the α and α' subunit mRNAs exhibit a higher degree of nucleotide conservation in their 3' noncoding regions than in their coding regions.

In this paper, we also present the partial DNA sequence for a gene encoding an α' subunit of the 7S soybean seed storage protein. Sequence comparison of the genomic clone and a homologous cDNA clone reveals the

presence of four small intervening sequences in the coding region of the α' subunit gene. The borders of the introns in this plant gene share extensive homology with the highly conserved 5' exon-intron border of vertebrate genes but not with the more variable 3' exon-intron border (9,10).

MATERIALS AND METHODS

Source of Recombinant cDNA Clones. The recombinant cDNA library was constructed in the Laboratory of Dr. L. Pollaro (University of Missouri, Columbia, Mo.) by using oligo(dT)12-18 to prime the first strand of double-stranded DNAs (11) complementary to poly(A)⁺ mRNAs, isolated from early maturation soybean seeds (stages R-1) (1). Secondary structure in the first DNA strand was used to prime the synthesis of the second DNA strand (12). The double stranded cDNAs were tailed with poly(dA) and ligated into the poly (dT)-tailed Hind III site of pBR322 (13,14). *E. coli* HB101 cells were transformed with the hybrid plasmids (15) and the resulting ampicillin-resistant, tetracycline-sensitive transformants were screened by transferring DNA fragments from each transformant to nitrocellulose filters (16) and hybridizing them with nick-translated DNA (17) from the α' subunit gene containing phage, ChA Gm α' 17.1. cDNA clones hybridizing with this probe were designated Gm α' 1, α' 2, etc. to indicate their homology with an α' subunit gene probe.

The 550 base pair long Gm α' 236 cDNA (Fig. 1) clone was initially described by Beachy *et al.* (3); further characterization of this clone is outlined in the accompanying paper (7). Intact Hind III sites, resulting from the addition of a Hind III linker to the 3' end of the double-stranded DNA insert and from the natural Hind III site at the 5' end of the cDNA insert, border the edges of the cloned Gm α' 236 insert (Fig. 1).

Isolation of Recombinant Phage containing Glycine max 7S Subunit gene. The glycine max library was constructed in Charon 4A lambda phage (18,19) by R. Nagao and R. Neugher (University of Georgia, Athens, Ga.). The primary screening was carried out by the plaque hybridization method of Woo (20) using as probe the 550 bp Hind III insert of Gm α' 236 cDNA labeled by nick-translation (17). Hybridizations were done at 42°C in 40% formamide, 0.60 M NaCl, 0.060 M Na citrate (4x SSC). The recombinant phage containing the α' subunit gene, ChA Gm α' 17.1, was purified on two successive CsCl block (0.14, 1.6) gradients. The phage were heated in 0.5% SDS, 5 mM EDTA (pH 8) for 15 min at 60°C and the DNA was isolated by phenol:chloroform:isoamyl alcohol extraction (50:50:1) followed by ethanol precipitation.

Construction of Plasmid Subclones from Phage with the α' Subunit gene. Plasmid subclones containing the 10.5 Kb and 1.6 Kb Eco RI-Pst I subfragments

[illegible]

sequencing of DNA, the method of Maxam and Gilbert (21) as modified by Smith and Gilbo (22) was used for sequencing DNA. Restriction fragments with 3' overhanging ends were labeled using reverse transcriptase and re-cut to

Results

The 550 base pair long cDNA, *ccat-2* (Fig. 1) used in further

Detailed Restriction Site Comparisons of the Class I and II cDNA Clones

close to the 5' end of the *gtaA* gene. The DNA was digested with *Sal*I and *Not*I, and the fragments were separated on agarose gels. The labeled DNA fragments were transferred to a nylon membrane and hybridized with the labeled 200 bp and 350 bp Hind III-*Pvu* II subfragments of *gtaA* α to γ . The cloned cDNA inserts were categorized into two major classes (Fig. 1) both

$\frac{1}{2} \text{C}_{10}\text{H}_{16} + \frac{1}{2} \text{C}_{10}\text{H}_{14} + \frac{1}{2} \text{C}_{10}\text{H}_{12} + \frac{1}{2} \text{C}_{10}\text{H}_{10} + \frac{1}{2} \text{C}_{10}\text{H}_8 + \frac{1}{2} \text{C}_{10}\text{H}_6 + \frac{1}{2} \text{C}_{10}\text{H}_4 + \frac{1}{2} \text{C}_{10}\text{H}_2 + \frac{1}{2} \text{C}_{10}\text{H} + \frac{1}{2} \text{C}_{10}\text{H}_0$

of which contain a conserved 0.47 Kb Hind III-Hinf I fragment complementary to the $\alpha'23b$ DNA. The cloned cDNAs in class I, $\alpha'16$ and $\alpha'17$, are distinguished by the presence of a 0.32 Kb Hind III-Hinf I fragment situated upstream from the 0.47 Kb Hind III-Hinf I fragment (Fig. 1) and a Pst I endonuclease site positioned 400 base pairs upstream from the Hind III restriction site. In contrast, the cloned cDNAs in class II have a Hinf I restriction site situated only 100 base pairs upstream from the 0.47 Kb Hind III-Hinf I fragment which shares sequence homology with $\alpha'23b$.

Isolation of a genomic DNA fragment containing a gene for the α' -Subunit.

a b c d e f g h i j k l m n o p q r s t u v w x y z
 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32

into the bacteriophage λ Charon 3A vector (19). The recombinant phage λ α 17.1 (Fig. 3) was identified by screening the library with the cloned cDNA (α 236 probe (Fig. 1). The region of homology between the α 17.1 gene and the α 236 cDNA was delineated by hybridization of phage DNA restriction fragments with the labeled 200 bp and 350 bp Hind III- α 111 restriction fragments of α 236 DNA (Fig. 1). Like the cDNA clones discussed earlier, this genomic DNA clone shared extensive homology with α 236 subunit mRNA specific region of α 236 as well as its α , α' , β -subunit mRNA common sequences. The extent and orientation of the α 17.1 gene were determined by hybridization of Charon α 17.1 DNA restriction fragments with both incomplete and full length oligo dT-primed reverse transcripts of mid-naturation seed mRNAs.

The restriction sites in the $\alpha_{17.1}$ gene were mapped with more precision after subcloning the 1.6 kb and the 10.5 kb Eco RI + *Pst* I fragments

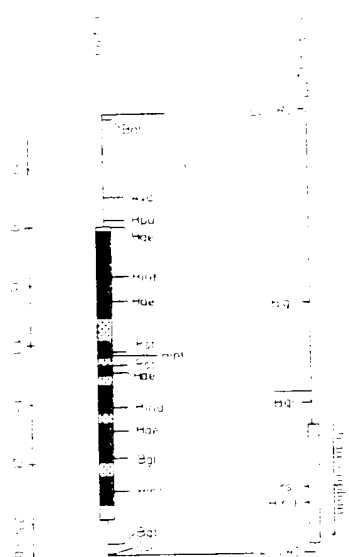


Figure 3. Restriction sites in the gene for an α -subunit of the 7S storage protein. The scale at the top of the figure indicates the size of the fragments in base pairs. The restriction sites for the enzymes used are indicated by arrows. The position of the gene is indicated by the open box. α and β subunits are indicated by open circles. The black dots represent the amino acid sequence of the α -subunit.

At the bottom, the *trnA* 17.1 gene is depicted in more detail. The boxes correspond to coding sequences; the open boxes corresponding with 5' and 3' noncoding sequences; the stippled boxes mark the intervening sequences found downstream from the first 4 sites in this gene. The existence of intervening sequences upstream from the first 4 sites has not yet been demonstrated. The origin of the scale below the gene is positioned at the site for transcription initiation (T.A. Scholten, unpublished results). The positions of the restriction sites in the gene and its flanking regions are based on the hybridization of genomic DNA fragments with full-length cDNAs complementary to total end-nucleation seed mRNAs, partial end-nucleotide (Lewy et al. and the first 3 bases) and DNA sequence analysis of the gene: BamI (Bcl II), HaeIII (Hae III), HinfI (Hinf I), AcaI (Aca II), HpaI (Hpa I), PstBI (Pst I) and HindIII (Hind III).

of *clbA* (*mgf* 17.1) into pBR322 (13). Inherent endonuclease sites having six base recognition sequences were determined relative to the Eco RI and Pst I sites by sizing the DNA fragments generated by the addition of Eco RI or Pst I to a two enzyme restriction digestion. Sites in the gene for enzymes with four base recognition sequences were determined by partial restriction enzyme digestion of the end-labeled 2.0 Kb Bgl II-Pst I and the 1.6 Kb Eco RI-Pst I fragments that contain the 5' and 3' halves of the *mgf* 17.1 gene, respectively. The length of the small internal Pst I fragment was deduced by comparison of the sequences of the Hind III-Pst I fragments in the cloned genomic DNA and cDNAs. The map sites and distances have been corroborated by sequence analysis of this gene. The information derived from these analyses has been compiled in the diagram at the bottom of Fig. 3. The positions and lengths of the 5' and 3' noncoding regions shown there are derived from

sequence analyses (this paper; M.A. Schuler, unpublished results).

Sequence analysis of the genomic DNA and cDNA clones complementary to α and α' subunit mRNAs. Although the amino acid sequence has not previously been determined for either the α or α' subunits, the high degree of similarity in amino acid composition of the α and α' subunits (1) and their proteolytic cleavage fragments (2) has suggested that the genes for these subunits are closely related. DNA sequence analysis was used to delineate the nucleotide homologies between the two classes of α and α' subunit cDNAs and to define the amino acids encoded by the α and α' subunit mRNAs. The DNAs cited in the comparison are: the genomic DNA clone, α g^d17.1, the representative class I cDNA clone, α cl^d16, the representative class II cDNA, α cl^d11, and the cDNA clone, α' cl^d36. On the basis of hybrid selection experiments (7; B.L. Ladin, unpublished results), α cl^d36 shares the greatest homology with α subunit mRNA and α g^d17.1 shares the most homology with α' subunit mRNA. The Ato I and H11 restriction sites (Fig. 2) which originally differentiated the (class I and) class II cDNAs, α cl^d16 and α cl^d11, also differentiate the α' 17.1 genomic DNA and α' cl^d36 cDNA. Whereas the restriction endonuclease sites in

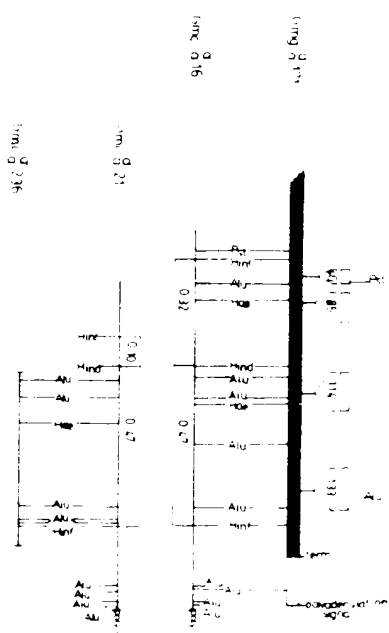


Figure 4. Summary of restriction map and sequence analysis of the cloned DNA for the α and α' subunits of 7S storage proteins. The restriction endonuclease sites shown in this figure have been determined by DNA sequence analysis and/or the restriction site mapping. The enzymatic cleavage sites are diagrammed in a palimpsest fashion which reflects the restriction site conservation in the α 17.1 cDNA and α' 16 cDNA clones for the α' -subunit and in the α 2.1 and α 2.3 cDNA clones for the α -subunit. In the diagram of the α 17.1 gene, the black box represents coding sequences and the open box, between the termination codon and the polyadenylation site, represents the 3' noncoding region of the gene transcript. The 40, 85, 115 and 133 nt. long intervening sequences are diagrammed above the gene. The restriction endonuclease abbreviations are the same as in Fig. 2.

amino acids
 10q11.1 gene nucleotide
 10q11.1 cDNA nucleotide
 10q11.1 cDNA nucleotide
 10q11.1 cDNA nucleotide

IVS 40

17.1 gene
 16 cDNA
 21 cDNA
 236 cDNA

IVS 85

17.1 gene
 16 cDNA
 21 cDNA
 236 cDNA

17.1 gene
 16 cDNA
 21 cDNA
 236 cDNA

17.1 gene
 16 cDNA
 21 cDNA
 236 cDNA

17.1 gene
 16 cDNA
 21 cDNA
 236 cDNA

17.1 gene
 16 cDNA
 21 cDNA
 236 cDNA

IVS 115

17.1 gene
 16 cDNA
 21 cDNA
 236 cDNA

17.1 gene
 16 cDNA
 21 cDNA
 236 cDNA

17.1 gene
 16 cDNA
 21 cDNA
 236 cDNA

17.1 gene
 16 cDNA
 21 cDNA
 236 cDNA

17.1 gene
 16 cDNA
 21 cDNA
 236 cDNA

IVS 132

17.1 gene
 16 cDNA
 21 cDNA
 236 cDNA

17.1 gene
 16 cDNA
 21 cDNA
 236 cDNA

17.1 gene
 16 cDNA
 21 cDNA
 236 cDNA

17.1 gene
 16 cDNA
 21 cDNA
 236 cDNA

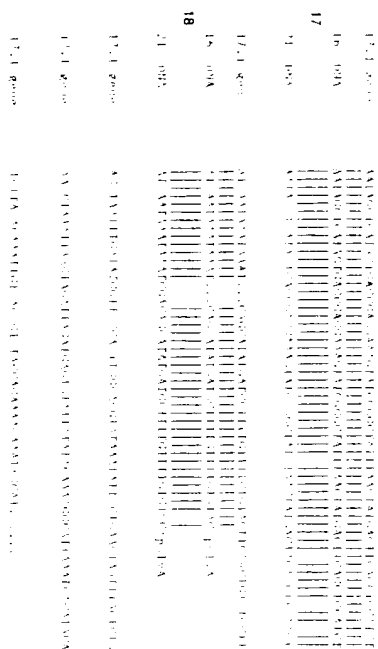


Figure 5. The nucleotide sequences of the cloned $\alpha^1 17.1$ genomic DNA and the $\alpha^1 16$, 21 and 23b cDNAs. The nucleotide sequence of the genomic α^1 -subunit clone is shown on line 2; the amino acids encoded by this sequence are shown on line 1. The positions of the 760 bp, 85 bp, 115 bp and 132 bp intervening sequences that interrupt the coding region of the genomic DNA are designated with arrows. Line 3 contains the nucleotide sequence of the $\alpha^1 16$ cDNA. The amino acid sequence of this cDNA is identical to that of the α^1 -subunit genomic DNA except at the positions designated on line 4. Vertical lines (|) mark the nucleotides that are identical in these two sequences. The nucleotide sequences of the $\alpha^1 21$ and the $\alpha^1 23b$ cloned cDNAs are shown on lines 6 and 7. The amino acid residues which differ from those found in the α^1 -subunit genomic DNA sequence are diagrammed above and below the nucleotide sequences on lines 6 and 7. The nucleotides conserved in both sets of paired DNA sequences are marked with solid vertical lines between the sequence sets. The amino acid variations occurring in the four sequences are highlighted with open boxes. Underlined nucleotides are designated by X.

genomic $\alpha^1 17.1$ clone resemble those in the cloned cDNA $\alpha^1 16$, the sites in the $\alpha^1 23b$ cDNA resemble those in the $\alpha^1 21$ cDNA. Comparative restriction site maps of the four DNA clones are shown in Fig. 4.

The four DNA sequences are diagrammed in Fig. 5 in a pairwise manner which indicates the close restriction site and nucleotide homologies existing between the $\alpha^1 17.1$ genomic and the $\alpha^1 16$ cDNA clones and between the $\alpha^1 21$ and $\alpha^1 23b$ cDNA clones. The DNA sequences include the last 912 nucleotides of the coding region in the α^1 -subunit DNAs and slightly fewer nucleotides in the c-subunit DNAs. They also include 132 nucleotides in the 3' noncoding region that extends from the translation termination signal to the polyadenylation signal of the mRNA. The sequence of the genomic $\alpha^1 17.1$ DNA clone contains 177 additional nucleotides which flank the 3' edge of the mRNA transcript.

The amino acid sequences derived from these nucleotide sequences indicate that the four cloned DNAs encode nearly identical mRNAs for the α or α^1 -subunits of the 7S soybean seed storage protein, because the only major

difference in the amino acid compositions of the α and α^1 -subunits occurs in the histidine content of the two proteins (3,7). The hybrid selection experiments (2,7; Bar, Jodan, unpublished results) provide the most convincing evidence presented here that the genomic $\alpha^1 17.1$ DNA and the $\alpha^1 16$ cDNA encode α^1 -subunits and that the $\alpha^1 21$ and the $\alpha^1 23b$ cDNAs encode c-subunits. Partial amino acid sequencing of the carboxyl-terminal fragment of the α^1 and α -subunits (H. Jodan, H. Jodan, and H. Jodan, manuscript in preparation) confirms that the genomic $\alpha^1 17.1$ DNA encodes an α^1 -subunit while the $\alpha^1 21$ cDNA encodes an α -subunit. The cloned cDNAs still retain both the α and α^1 immunofluorescence because of the nucleotide similarities between the α and α^1 -subunit cDNAs.

The coding regions of the closely related $\alpha^1 17.1$ genomic DNA and the $\alpha^1 16$ cDNA contain 3/604 base mismatches (0.5%); a similar region in the $\alpha^1 21$ cDNA and the $\alpha^1 23b$ cDNA possesses 9/300 base mismatches (3%). Intercomparison of the α and α^1 -subunit DNAs shows that approximately 50/744 base mismatches (7%) occur between the coding regions of these DNAs. Although a high degree of nucleotide homology exists among all four DNAs, it is evident that the highest degree of sequence conservation occurs within the classes of paired DNA sequences. Nucleotide differences, insertions or deletions similar to those which distinguish the Class I and Class II DNA clones occur in the coding and 3' noncoding sequences of all other α and α^1 -subunit cDNA clones which have been sequenced, with the exception of one in which the poly(A) tract (lines 15 nucleotides upstream from the poly(A) tract) found in the other cDNAs (data not shown).

The high degree of codine nucleotide conservation provides for extraordinary conservation in the amino acid residues encoded by the α and α^1 -subunit DNAs since many of the nucleotide differences occur in the third base of the amino acid codons. Within each class of paired DNAs, either two or five amino acid differences occur in the last 175 amino acids of the encoded proteins. In contrast, when the last 175 amino acid residues of the α and α^1 -subunits are compared, 27 amino acid differences are evident. The greatest concentration of amino acid differences in the proteins encoded by the α and α^1 -subunits cDNAs occurs in the region 31 to 46 amino acids before the carboxyl-terminal of these proteins. This region of the amino acid variation corresponds with the region in the $\alpha^1 21$ cDNA which contains several closely spaced Alu I restriction sites (Fig. 2) and which lies at the right edge of the coding region conserved in the mRNAs for the 7S storage proteins and other seed proteins (7).

Interpretation of Conservation and Secondary Structure of the 3' Noncoding Regions of the Class I and II Cloned cDNAs

The nucleotide homologies between the 3' noncoding regions of α for the Class I α -subunit cDNA and α' 21, the Class II α -subunit cDNA, are more extensive than those found in some sections of the coding sequences of these cloned cDNAs; only 6/132 base mismatches occur in the 3' noncoding sequences of the α and α' -subunit cDNAs, whereas 13/161 mismatches occur in the region upstream from the termination codon in which the most intensive amino acid variation occurs. In addition, different 4 bp insertion or deletion sequences occur in the 3' noncoding sequences of the α and α' -subunit cDNAs. The nucleotide conservation in the 3' noncoding regions of these soybean cDNAs is highly unusual and implies that, since the implication of an ancestral α' -subunit gene sequence, functional constraints have prevented the divergence of the 3' noncoding sequences. Because the secondary structure of the mRNA transcripts may have constrained the evolution of the α and α' -subunit gene sequences, the 3' noncoding regions were analyzed for secondary RNA structures according to the base-pairing rules of Tinoco et al. (25). In the most stable conformation derived for the mRNA transcripts (Fig. 6), a large fraction of the conserved 3' noncoding nucleotides form double-stranded structures. Most of the nucleotide differences in the noncoding regions of the α and α' -subunit mRNAs lie outside the double-stranded regions. The entire double polynucleotidation signal (AAUAAAUAA) (25) found in these cloned cDNAs resides in a single stranded RNA loop.

The Intervening Sequences of the α' -Subunit Gene. Although DNA restriction site analyses of the regions downstream from the Pst I sites in the cloned α' 17.1 genomic DNA and the α' 16 cDNA detected little difference in the sizes of the restriction fragments of the genomic DNA and cDNA, DNA sequence analysis of this region demonstrated that four introns are present in the genomic DNA clone. The introns are 85, 115, 132 and approximately 40 nucleotides in length and are positioned as diagrammed in Fig. 5. The sequences of the 85, 115 and 132 nucleotide introns and the coding sequences which they interrupt are shown in Fig. 6. The smallest intron has not been sequenced. The 5' exon-intron junction of the 115 and 132 nt. introns 5'-GAGGTAAG-3' and 5'-GAGGTAAT-3', contain the core pentanucleotide AGGTA found in the consensus sequence derived from other eukaryotic exon-intron junctions (9,10).

The 3' junctions of the three introns in this α' -subunit gene exhibit somewhat less homology with the consensus sequence for the 3' intron-exon junction derived for vertebrate genes (9,10). Only the AG dinucleotide at positions -1 and -2 in the 3' splice junction is similar in the plant and

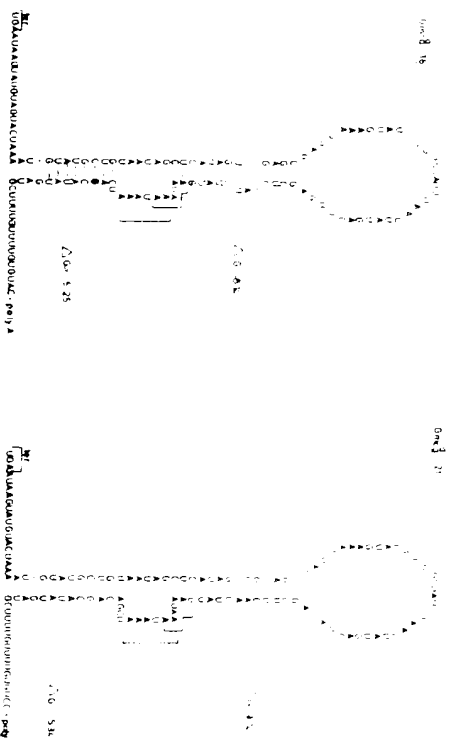


Figure 6. Potential secondary structures of the 3' noncoding sequences in the α and α' -subunit mRNAs. The RNA sequences of the last 100 coding nucleotides and the 3' noncoding nucleotides of α 16, a Class I cDNA, and α' 21, a Class II cDNA, were analyzed for potential secondary structures according to the rules of Tinoco et al. (25). The most stable structures derived from this analysis are shown here; horizontal lines connect the base paired nucleotides. Free energy potentials are shown at the right of each base paired structure. The repetitive AAUAA polynucleotidation signals (25) are bracketed.

vertebrate gene introns. The polypyrimidine-rich region, which appears to constitute part of the splice signal of the 3' intron border (10), stretches 24 nucleotides in front of the splice points in the 85 and 115 nt. intervening sequences. Although a limited number of nucleotide positions are conserved in the 3' border sequences of the 85, 115 and 132 nucleotide introns, the homologies that are found in these three soybean gene intron junctions do not occur in the three introns of the phaseolin storage protein gene (26). Purine di-, tri- and tetranucleotides occur with a high frequency in nucleotide positions -3 to -24 in the intron borders of the glycine and phaseolin storage protein genes. This contrasts sharply with the 3' border sequences found in a multitude of vertebrate genes (10).

DISCUSSION

It is becoming increasingly evident that many eukaryotic cells which must produce abundant proteins in a relatively short period of time do so by regulating the expression of families of closely related genes. In this paper, we have characterized one genomic DNA clone and numerous cloned cDNAs which code for the α and α' -subunits of the 7S seed storage protein in

Schulter, unpublished results). These differences may reflect the fact that the two polyadenylation signals are not equally accessible to mRNA processing enzymes, possibly because of secondary RNA structures such as those shown in Fig. 9. The sequences immediately upstream from the poly(A) tails in the short and long versions of the 3' noncoding sequences of the α and α' -subunit mRNAs do not contain homologous nucleotides. This suggests that the polyadenylation signal alone is sufficient for positioning the RNA processing enzymes.

Four small intervening sequences have been detected in the genomic α' -subunit DNA clone through comparison of the genomic DNA and cDNA sequences. The 3' exon-intron junctions of these introns match the intron consensus boundary defined for the vertebrate genes (9,10). In contrast, the 3' intron-exon boundaries in this plant gene share little sequence homology with the 3' intron-exon border sequences found in vertebrate genes; only the AG dinucleotide at the 3' splice point of the intron is conserved. Unlike the introns in vertebrate genes, but like the introns of the closely related phaseolin gene (20) Schulter et al., manuscript in preparation), the soybean α' -subunit gene has an extremely high concentration of purine bases in the 24 nucleotide sequence preceding the 3' splice junction. If these nucleotides are important in determining the proper splice point in a precursor RNA, as has been previously speculated (10), then the RNA or protein moiety which recognizes this sequence differs in plants and animals.

The homologies in the cDNAs encoding the α and α' -subunits indicate that the genes for the α and α' -subunits have evolved from a common ancestral sequence. The nucleotide similarities within each gene family indicate that duplicate copies of a primordial α/α' gene sequence diverged substantially from one another and produced distinct α and α' -subunit gene sequences before further duplication events occurred. The individual ancestral α and α' -subunit gene sequences independently underwent a series of duplications subsequent to this initial divergence. The genes produced in this final series of duplications have continued to diverge from one another to produce the members of the present α and α' -subunit gene families. In many respects, the evolution of the α and α' -subunit gene families parallels the evolution of the well-studied α and β -globin genes (review, 23), although, the nucleotide sequences of the 7S α and α' -subunit gene families are more closely related than those of the α and β -globin genes.

ACKNOWLEDGEMENTS

The authors gratefully thank Dr. J. Pollaco and Dr. P. Frey for the cloned soybean cDNA library, Drs. R. Dechard and R. Dayan for the recombinant glycine max genomic library and Dr. M.H. Bolin and J. J. Giese for the preliminary mapping of the recombinant phage. We are especially grateful to Dr. L.H. Boxman for help with the computer analysis of the secondary RNA structure. This work was supported by grants from the Department of Energy, DE-AC02-81 ER10088, the United States Department of Agriculture, SEA-CR60, 58-2094-1-1-705-01, and the National Science Foundation, PCM-791763.

To whom correspondence should be addressed.

REFERENCES

1. Deicke, D.H., Chen, J., and Beachy, R.H. (1981) *Planta* 153, 130-139.
2. Thamb, V.H., and Shibasaki, K. (1976) *Biochim. Biophys. Acta*, 439, 376-388.
3. Beachy, R.H., Jarvis, N.P., and Barton, K.A. (1981) *J. Mol. Appl. Genet.* 1, 19-27.
4. Holwachs, L.P. (1981) Ph.D. Dissertation (Cornell University, Ithaca, N.Y.)
5. Thamb, V.H., and Shibasaki, K. (1977) *Biochim. Biophys. Acta* 490, 370-384.
6. Beachy, R.H., Barton, K.A., Thompson, J.F., and Madison, J.T. (1980) *Plant Physiol.* 65, 990-994.
7. Schulter, H.A., Ladin, B.F., Pollaco, J.C., Frey, G., and Beachy, R.H. (1982) *Nucleic Acids Res.* accompanying paper.
8. Goldberg, E.B., Hoshok, G., Ditta, G.S., and Breidenbach, R.W. (1981) *Dev. Biol.* 83, 218-231.
9. Breidenbach, R., Benoit, C., O'Hare, K., Cannon, F., and Chabon, P. (1978) *Proc. Natl. Acad. Sci. USA* 75, 4853-4857.
10. Leiner, M.K., Boyle, J.A., Mount, S.M., Molin, S.A., and Soltz, J.A. (1980) *Nature* 283, 220-224.
11. Hyatt, J.C., and Spiegelman, S. (1978) *Proc. Natl. Acad. Sci. USA* 75, 5329-5333.
12. Weiland, H., Bruschke, C., and Felix, G. (1979) *Nucleic Acids Res.* 6, 2707-2715.
13. Bolivar, G., Rodriguez, R.L., Greene, P.L., Betlach, M.C., Mesner, H.L., and Boyer, H.W. (1977) *Gene* 2, 95-113.
14. Clarke, L., and Carbon, L. (1975) *Proc. Natl. Acad. Sci. USA* 72, 4301-4305.
15. Messink, P.C., Finegan, D.L., Donelson, J.E., and Hogness, D.S. (1974) *Cell* 3, 315-325.
16. Wall, G.H., Stern, M., and Stark, G.R. (1979) *Proc. Natl. Acad. Sci. USA* 76, 3683-3687.
17. Maniatis, T., Jeffrey, A., and Kleid, D.G. (1975) *Proc. Natl. Acad. Sci. USA* 72, 1184-1188.
18. Blattner, F.R., Williams, B.G., Blechl, A.E., Thompson, K.D., Faber, H.E., Furlong, L.A., Grunwald, D.J., Kleier, D.O., Moore, D.B., Schumm, J.W., Sheldon, K.J., and Smithies, O. (1977) *Science* 196, 161-169.
19. Maniatis, T., Harbison, R.C., Lacy, E., Lauer, J., O'Connell, C., Ono, D., Sim, G.K., and Efstratiadis, A. (1978) *Cell* 15, 687-701.
20. Igo, S.L.C. (1979) in *Methods in Enzymology*, Wu, R., Ed. Vol. 68, pp. 349-395, Academic Press, New York.
21. Maxam, A.M., and Gilbert, W. (1980) in *Methods in Enzymology*, Grossman, L., and Moldave, K., Eds. Vol. 65, pp. 499-560, Academic Press, New York.
22. Smith, D.R., and Galvo, J.H. (1980) *Nucleic Acids Res.* 8, 2255-2274.
23. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, G., Sprints, R.A., Telford, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L.,

- Bruch, A.C., Smithies, O., Brattley, F.E., Shoofers, C.C., and Proudfoot, L.J. (1980) Cell 21, 693-698.
24. Finney, L., Borer, F.B., Bengler, B., Levine, H.D., Hildebrandt, U.C., Grotzinger, D.H., and Gellera, L. (1979) Nature Rev Biol. 56, 369-64.
25. Proudfoot, L.J., and Brownlee, G.B. (1976) Nature 263, 241-243.
26. Song, S.H., Shihone, T., and Hall, T.G. (1981) Nature 289, 37-41.
27. Chen, P.Y., and Pasquini, G.D. (1977) Biochem. 13, 244-249.
28. Chen, P.Y., and Pasquini, G.D. (1977) J. Mol. Biol. 115, 185-175.
29. Gantier, L., Ouchetto, D.J., and Robson, K. (1978) J. Mol. Biol. 120, 97-100.
30. Forbush, L.A., Bond, B.L., Hershey, H.D., Hixson, F.S., and Davidson, H. (1981) Cell 24, 107-116.
31. Scholter, D.A., Hershov, P., and Keller, B.B., manuscript submitted.
32. Hartin, S.L., Zuber, P.A., Davidson, W.S., Wilson, A.C., and Kan, Y.M. (1981) Cell 25, 737-741.
33. Skolnik, L., and Frey, C., and Roopman, F. (1980) Nucleic Acids Res. 8, 343-355.

Structural sequences are conserved in the genes coding for the α , α' and β -subunits of the soybean 7S seed storage protein

May A Schuler¹, Beth E. Adam², Joseph C. Pollock¹, Gregory Feyer¹, and Roger N. Beachy¹

¹Plant Biology Program, Department of Biology, Washington University, St. Louis, MO 63130, and
²Biochemistry Department, University of Missouri, Columbia, MO 65212, USA

Received 25 June 1982; Revised 28 September 1982; Accepted 11 October 1982

ABSTRACT

Cloned mRNAs encoding four different proteins have been isolated from recombinant cDNA libraries constructed with glycine max seed mRNAs. Two cloned mRNAs code for the α and α' -subunits of the 7S seed storage protein (conglycinin). The other cloned cDNAs code for proteins which are synthesized *in vitro* as 68,000 d., 60,000 d., or 53,000 d. polypeptides. Hybrid selection experiments indicate that, under low stringency hybridization conditions, all four cDNAs hybridize with mRNAs for the α and α' -subunits and the 68,000 d., 60,000 d., and 53,000 d. *in vitro* translation products. Within three of the mRNAs, there is a conserved sequence of 155 nucleotides which is responsible for this hybridization. The conserved nucleotides in the α and α' -subunit cDNAs and the 68,000 d. polypeptide cDNAs span both coding and noncoding sequences. The differences in the coding nucleotides outside the conserved region are extensive. This suggests that selective pressure to maintain the 155 conserved nucleotides has been influenced by the structure of the seed mRNA. RNA blot hybridizations demonstrate that mRNA encoding the other major subunit (β) of the 7S seed storage protein also shares sequence homology with the conserved 155 nucleotide sequence of the α and α' -subunit mRNAs, but not with other coding sequences.

INTRODUCTION

Literature on the expression of the genes for the legume seed storage proteins has been accumulating rapidly. The studies deal with a variety of legumes, including *Glycine max* (soybean), *Phaseolus vulgaris* (French garden bean) and *Pisum sativum* (Garden pea), and include characterization of storage protein complexes by sucrose gradient fractionation (1,2), the storage protein subunits by peptide mapping (3,4,5) and characterization of the mRNAs for the storage proteins by *in vitro* translation assays (3,4,6,7,8,9). From this work, two major classes of storage proteins referred to as the legumins (11S sedimentation coefficient) and the vicilins (7S sedimentation coefficient) (2) have been identified in most legumes. Both the 7S and 11S classes of storage proteins contain a number of closely related major subunits (3,5,10). The similarities in the subunit organization and the amino acid compositions of the various legumin and vicilin

GENE 74:433-443

Organization of the sunflower 11S storage protein gene family

(Legumin/globulin seed proteins; nucleotide sequence; divergent gene families; *Helianthus annuus*; helianthinin)

Raymond A. Vonder Haar, Randy D. Allen*, Elizabeth A. Cohen*, Craig L. Nessler and Terry L. Thomas

Biology Department, Texas A&M University, College Station, TX 77843 (U.S.A.)

Received 13 April 1988

Accepted 13 August 1988

Received by publisher 19 September 1988

SUMMARY

We have isolated and characterized genes encoding the sunflower 11S globulin seed storage proteins, collectively termed helianthinin. One gene, designated *HaG3*, has a primary transcription unit of about 1750 nucleotides including two short intervening sequences. The predicted precursor polypeptide from *HaG3* is 493 amino acids long, is rich in glutamine and other nitrogen-rich amino acids and includes the amino acid sequence NGVEETICS. This sequence is highly conserved among 11S seed storage proteins and is involved in the proteolytic processing of these polypeptides. Additional helianthinin sequences are conserved among other seed storage protein genes. Analysis of various cDNA and genomic sequences indicates helianthinins are encoded by a small gene family that includes a minimum of two divergent subfamilies.

INTRODUCTION

Like embryos of other oilseed plants, sunflower embryos accumulate and store large quantities of lipid and protein. These stored materials are utilized by the seedling following germination and, in addition, are of immense agronomic importance. The organization and expression of genes encoding seed

storage proteins has been investigated in a number of plant species, including both dicots and monocots (reviewed by Shotwell and Larkins, 1988). In all cases, the accumulation of storage proteins during seed development and maturation requires the highly regulated expression of genes encoding these proteins. Substantial post-translation modifications and targeting to appropriate subcellular compartments

Correspondence to: Dr. T.L. Thomas, Biology Department, Texas A&M University, College Station, TX 77843 (U.S.A.), Tel. (409) 845-0184; Fax (409) 845-2891.

* Present addresses: (R.D.A.) Department of Biology, Washington University, St. Louis, MO 63130 (U.S.A.), Tel. (314) 889-6883; (E.A.C.) Department of Genetics, University of Georgia, Athens, GA 30602 (U.S.A.), (404) 542-1444.

Abbreviations: aa, amino acid(s); bp, base pair(s); 2D, two dimensional; DPF, days post-flowering; Denhardt's solution, 0.02% bovine serum albumin, 0.02% Ficoll and 0.02% polyvinylpyrrolidone; ER, endoplasmic reticulum; nt, nucleotide(s); ORF, open reading frame; PAGE, polyacrylamide gel electrophoresis; SDS, sodium dodecylsulfate; SET, 0.15 M NaCl, 0.02 M Tris-HCl, 0.002 M EDTA (pH 8.0). For nucleotide sequences, H = A, C or T; M = A or C.

are also necessary. Consequently, these genes provide an excellent opportunity for analysis of the molecular mechanisms controlling many aspects of ontogenic gene expression in plants.

Sunflower seed proteins include the water soluble 2S albumins and the salt soluble 11S globulins. The sequence and expression of albumin structural genes has been described (Allen et al., 1987a, 1987b). The sunflower 11S storage protein, designated helianthinin, is structurally similar to legumin-like seed proteins of other plant species and is represented in planta by an approximately 300-kDa hexameric holoprotein (reviewed by Shotwell and Larkins, 1988). Each subunit of the holoprotein consists of a larger, acidic (α) polypeptide (30–40 kDa) and a smaller, basic (β) polypeptide (23–27 kDa) linked by disulfide bonds. The α and β polypeptides of legumin-like proteins such as the helianthinins are generated proteolytically from larger precursor polypeptides that are synthesized on the rough ER. An NH_2 terminal signal peptide targets the nascent polypeptide to the lumen of the rough ER, where it is removed. The 11S precursors assemble into trimers in the ER and are then transported to the vacuole through the Golgi. Once in the vacuole, 11S precursors are cleaved into disulfide-linked α and β polypeptides. The trimers then assemble into hexamers, and following additional protein accumulation, the vacuole subdivides to form protein bodies characteristic of many plant seeds (Higgins, 1934).

The cloning and expression of helianthinin mRNAs has been described (Allen et al., 1985; Allen et al., 1987b). Synthesis of helianthinin mRNAs and precursor polypeptides is tightly regulated during sunflower embryogenesis. Helianthinin α and β subunits first appear about 7 DPF, two days after the albumin seed proteins appear (Cohen, 1986), and like the albumins, these polypeptides continue to accumulate through much of sunflower seed development. Helianthinin mRNAs are also detected 7 DPF; these transcripts accumulate and disappear with kinetics similar to those observed for albumin mRNA (Allen et al., 1987b).

In this paper, we describe the isolation and characterization of genes encoding helianthinin in sunflower. Sequence and S1 nuclease analysis of one gene, designated *HaG3*, defined a primary transcription unit of about 1750 nt, including two short intervening sequences. The helianthinin polypeptide

predicted from the nucleotide sequence of *HaG3* shares significant, functional sequence homologies with other 11S seed storage proteins. Analysis of cDNA and genomic DNA sequences indicate helianthinins are encoded by a small gene family that includes at least two divergent subfamilies. Sequences located 5' of the *HaG3* transcription unit are conserved among other seed storage protein genes.

MATERIALS AND METHODS

(a) Materials

Sunflower seeds (*Helianthus annuus* L. cv. Giant Grey Stripe, Northrup King Seed Co., Minneapolis, MN) were obtained commercially. Embryos from field-grown plants were dissected from achenes at the indicated times, frozen in liquid nitrogen and stored at -80°C .

(b) Isolation and labeling of nucleic acids

Bacteriophage and plasmid DNAs were prepared by standard methods (Maniatis et al., 1982). Total and poly(A)⁺ RNA from leaves and staged sunflower embryos were prepared as described by Allen et al. (1985). Radiolabeled hybridization probes for genomic library screening, phage recombinant mapping and genomic DNA blots were prepared by nick translating a 1.1-kb *EcoRI* fragment prepared from the cDNA recombinant, *Ha2* (Allen et al., 1987a; Allen, 1986).

(c) Plaque hybridization

Construction of a sunflower genomic library in the bacteriophage λ vector EMBL3 (Frishauf et al., 1983) has been described (Allen et al., 1987a). The library was screened for helianthinin phage recombinants by hybridization with nick translated *Ha2* cDNA probes (Benton and Davis, 1977). Filters were prehybridized 4 h and hybridized 15–18 h at 67°C in $4 \times \text{SET}$, $5 \times \text{Denhardt}$, 0.2% SDS, 100 $\mu\text{g/ml}$ denature calf thymus DNA, 50 $\mu\text{g/ml}$ poly(A) and 10 $\mu\text{g/ml}$ poly(C). Filters were washed successively at 60°C in $4 \times$, $2 \times$, and $1 \times \text{SET}$ containing 0.025 M phosphate buffer and 0.2% SDS.

for each, air-dried and autoradiographed. Positive recombinants were plaque-purified and restriction-mapped by standard procedures (Maniatis et al., 1982).

(d) Nucleotide sequence analysis

HaG3 DNA was sequenced by the dideoxynucleotide chain termination method (Sanger et al., 1977) after ligation into M13mp18 and M13mp19 and transfection into JM101 (Messing et al., 1983). Single-stranded recombinant phage DNA was processed and sequenced as described (Sanger et al., 1980). Additional overlapping T4 polymerase dele-

tions of selected recombinants were prepared and sequenced as described by Dale et al. (1985). The complete sequence of *HaG3* was assembled from these overlapping clones. Computer analyses were done on a DEC MicroVax using the University of Wisconsin Genetics Computer Group (UWGGC) Sequence Analysis Software (Version 5.0; Devereux et al., 1984).

(e) Transcription analysis

Nuclease mapping of the transcriptional start point of *HaG3* was done as described by Favaloro et al. (1980) using a 446-bp *XhoI-DraI* fragment (see

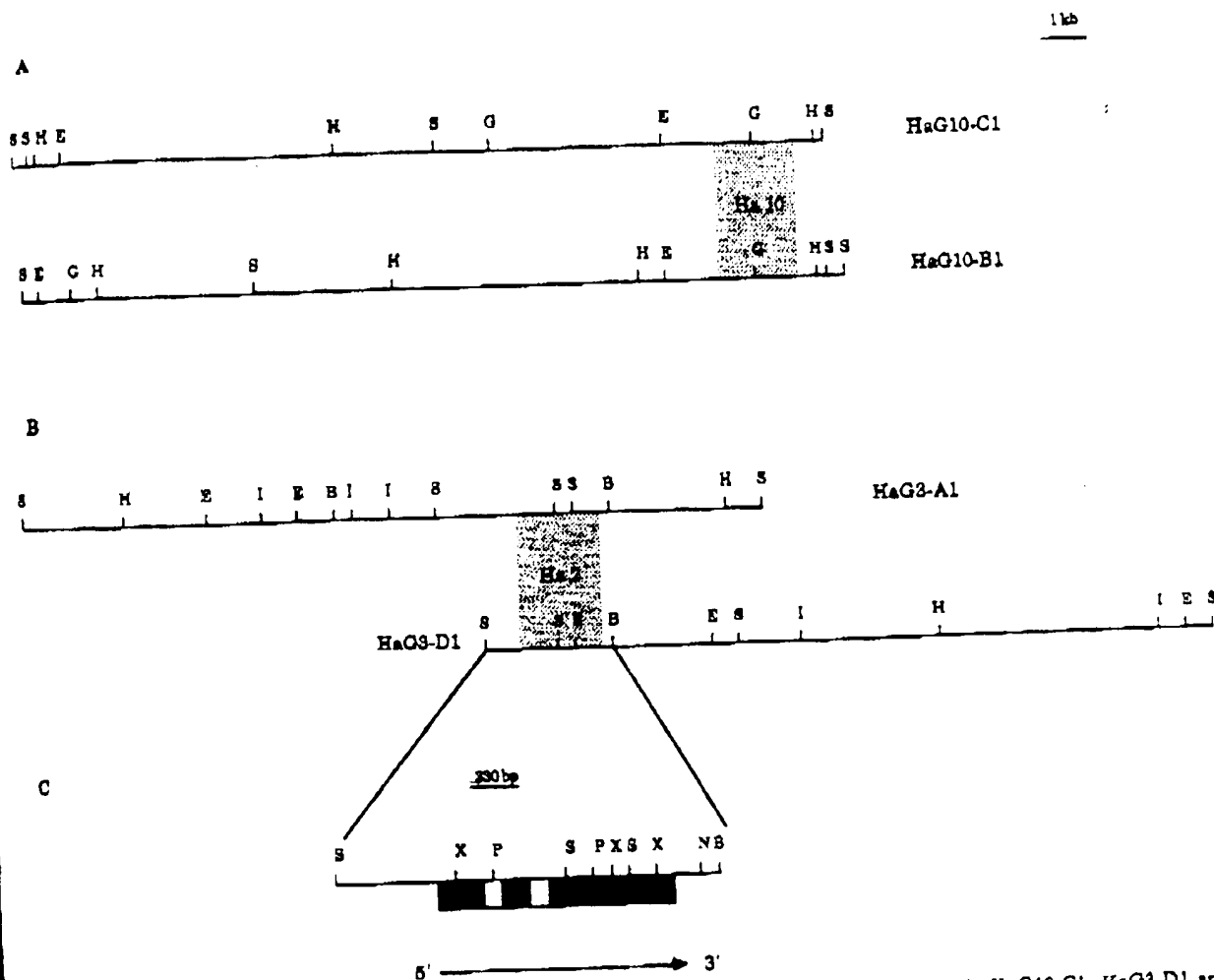


Fig. 1. Physical maps of sunflower helianthinin genes. Panels A and B: Restriction maps of *HaG10-B1*, *HaG10-C1*, *HaG3-D1* and *HaG3-A1*. Shaded areas indicate regions which hybridize with helianthinin cDNAs, *Ha2* and *Ha10*. Panel C: Detailed map of 2.8-kb *HaG3-D1*. Shaded area indicates the organization of the helianthinin transcription unit in *HaG3-D1* is shown. Dark *Sall-BglII* fragment of *HaG3-D1* that was sequenced. Organization of the helianthinin transcription unit in *HaG3-D1* is shown. Dark boxes indicate exons; open boxes indicate introns. B, *BglII*; E, *EcoRI*; G, *BglI*; H, *HindIII*; I, *BamHI*; N, *NcoI*; P, *PstI*; S, *Sall*; X, *XhoI*.

Sal I
1 GTCGACTATCTTAGTTAATCAAATAAAATTTATTTTGATTGCTTTTGTTAATGTAATTTCTCTAGTTTAA
71 AGTCGATCTGTATTTATATAATATTAGTAATTTTTATTAACATCAATACATGCTTCAGCTTTTGTGTTA
141 GTCTTCGTTTTTTATATGGTTTTATCAGCGGTGTGGTGTACGATGACGATTATTTAAATAATGACCGACT
211 TCTTGGTTGTTACTTATTGATGTACGAAGCTGAGATGTAACGAACCGAACACATATAAATAACATTTTGG
281 ATAAGATTACGACTTTATTTATCGGTGCCATGAAATTTGGAAGACTTGGGTAAAGACACAACCACATAT
351 AATGTGATGGTAAATAGCATTACAACTAATGTTAATCTTTTGTTACAAATGTTGTTAACTAGGCTTGAT
421 ATGTAAAATTTTAAAGACTATATGGTGTCTTACGGTTTTACATCTAGTAAGAGATTAAAAAAAAAAAA
491 AAGCAAGCAAAGTAAGTGTAAGAGAGTAAGAGAAATGTAGCCATCATATGGCTGATTGTTTCATCACCAT
561 CCCATTTATACTTATCATCTTGATGATGCATATAGACAACACACTACTTATACAGATGTAGCATGTCTC
631 AGCTGCAAAATGGTGATCTTCTCCTGGCATAACCTCTTAGATGTCACTTCCTCCTTGATCTTCTTCCACTA
701 TAAACCAGCTAGTTCACAACACCTATTACCCACATCACATCCCATTCCACTTAACAATGGCATCCAAAC
M A S K A
771 CAACTTTTGCTCTTAGCTTTTACCCCTTCTCTTTGCCACTTGCATTGCCCGCCACCAGCAACGGCAACAGCA
T L L L A F T L L F A T C L A R H Q Q R Q Q Q
Xho I
841 ACAGAACCAAGTGCCAGCTTCAAAACATCGAGGCGCTCGAGCCCATCGAAGTTATCCAAGCTGAAGCCGCT
Q N Q C Q L Q N I E A L E P I E V I Q A E A G
911 GTGACCGAAATTTGGGACGCTATGACCAACAGTTCAGTGTCCGTGGTTCGATTTTATTTCGACACCGGAT
V T E I W D A Y D Q Q F Q C A W S I L F D T G F
981 TCAACCTGGTGGCCTTCTCTTGCCCTTCTAGGTCAACACCCCTATTTTGGCCTTCGTGAGAGAGGTATA
N L V A F S C L P T S T P L F W P S S R E <----
Pst I
1051 CACATAAATAAATATTTTAAGAGTGGCAAATTAAGTTTAAAAATAATAATCTAACTGCAGTGTTTTGGC
----- Intron 1 -----
1121 ATGTTTAAAGGTAGGGGTATTCAAGGGGTATATTGCCGGGATGCCCCAGAACCTATGAATATTCCGAGG
----->G V I L P C C R R T Y E Y S Q E
1191 ACCAACAGTTTTCCGGTGAGGGTGCGCCGAGAGGAGGAGGAGGGCAGATTACGACCGTCAATCAGAAA
Q Q F S C E G G R R C G G E G T F R T V I R K
1261 GTTAGAGAACTTAAAGCAGGGTGACGTGGTTGCCATCCCCACCGGAACAGTCACTGGCTTACAAGGAC
L E N L K E G D V V A I P T G T A H W L H N D
1331 GGCAACACAGAACTTGTGGTCTGCTCTTGGATACTCAGAACCATGAGAACCAGCTTGACGAAAACCAA
G N T E L V V V F L D T Q Y H E N Q L D E N Q R
1401 GGGTAACATATATACTCTAAAAAATACTCCATTTTAAACCTAAATATATATACTGAACATAAACTT
<----- Intron 2 -----
1471 GTAACGTTTCAGAGATTCTTCTTAGCCGGAACCCCTCAAGCTCAAGCTCAAAGCCAGCAGCAACAACAA
----->R F F L A G N P Q A Q A Q S Q Q Q Q Q R

Fig. 2, nt position 433 to 879) asymmetrically labeled at the 5' terminus of the *Xho*I site. Total embryo RNA was used. The only differences in method were that the hybridizations were carried out for 6–8 h and 10 units of S1 nuclease were used per reaction. Reaction products were analyzed on polyacrylamide-urea gels. The 446-bp *Xho*I-*Dra*I fragment was subjected to Maxam-Gilbert sequencing reactions (Maxam and Gilbert, 1980) which were then used as length markers in the S1 protection experiments.

RESULTS AND DISCUSSION

(a) Isolation and characterization of helianthinin genes

A cDNA recombinant representing helianthinin mRNA was used to screen a sunflower genomic DNA library constructed in the bacteriophage λ vector, EMBL3 (Frishauf et al., 1983). Multiple bacteriophage λ recombinants representing the helianthinin gene family were recovered in these screens. Further analysis of these recombinants by hybridization with the divergent helianthinin cDNAs, *Ha2* and *Ha10* (Allen et al., 1987b), defined two divergent subfamilies that encode helianthinin in sunflower embryos (Fig. 1, A and B). Two bacteriophage λ recombinants, *HaG10-B1* and *HaG10-C1*, hybridize primarily to *Ha10*; under less stringent hybridization criteria ($6 \times$ SET, 55°C), these recombinants cross-hybridized weakly with *Ha2* (data not shown). Conversely, *HaG3-D1* and *HaG3-A1* were more similar to *Ha2* than to *Ha10*. Additional sequence data presented below confirms these sequence relationships.

(b) Sequence of the *HaG3* helianthinin gene

A 2.8-kb region of the genomic recombinant, *HaG3-D1*, bounded by *Bgl*II and *Sa*I sites (Fig. 1C), was sequenced to determine the precise organization of a representative sunflower legumin-like seed storage protein gene (Fig. 2). Three exons separated by two very short introns (99 and 79 bp) were identified by comparison to the amino acid sequence predicted from the helianthinin cDNA,

Ha2 (Allen, 1986). Intron/exon boundaries were assigned based on ORF discontinuities at each junction, on the colinearity of *HaG3* and *Ha2* on either side of each intron and on the presence of consensus splice junctions (Mount, 1982). The locations of the three exons and two introns in *HaG3-D1* are schematically shown in Fig. 1C; the precise sequence locations are displayed in Fig. 2.

The introns in the *HaG3* transcription unit differ in number and location from those observed for the prototypical *legA* gene of pea (Lycett et al., 1984). The *legA* gene has three introns at aa positions 95, 179 and 388 (henceforth referred to as I1, I2 and I3). The *HaG3* legumin gene has two introns at approximately the same positions as I1 and I2; I3, however, is missing from the sunflower gene. The pea *legJ/K* genes (Gatehouse et al., 1988) and the *Vicia faba* *LeB4* gene (Bäumlein et al., 1986) each contain two introns; in these genes however, I2 and I3 remain and I1 is absent. Interestingly, two divergent *Arabidopsis* legumin genes contain all three introns in approximately the same relative position as previously noted for the pea *legA* gene (Pang et al., 1988).

The *HaG3* transcription unit was mapped by S1 nuclease protection (data not shown). The transcriptional start point is located at nt position 726 (Fig. 2), 32 nt upstream from the translational initiation site. Consensus sequence elements typical of RNA polymerase II transcription units in the regions surrounding the legumin transcription unit are underlined in Fig. 2. These include a CAAT homology at nt position 635 and a TATA homology at position 699, both 5' of the transcriptional start point. A consensus polyadenylation signal, AATAAA, is located 37 nt 3' of the stop codon.

Sequence elements located 5' of the *HaG3* transcription unit are shared with upstream sequence elements associated with other storage protein genes. Particularly noteworthy is the conservation of an element of the legumin (*leg*) box, a phylogenetically conserved sequence located approximately 100 nt upstream from several genes encoding legumin and legumin-like seed proteins (Bäumlein et al., 1986). Although the complete *leg* box is not conserved in *HaG3*, three elements that differ from the sequence AGAATGTC by only one nt are located between 50 and 210 nt upstream of the *HaG3* cap (indicated by a in Fig. 2). In addition to elements

the leg box, the consensus sequence, HAACAC-AM, characteristic of most seed protein genes (Goldberg, 1986) is present at position 598 in Fig. 2 (indicated by b). Despite the conservation of sequence and location of the legumin box elements and the CACA motif in *HaG3*, the functional significance of these conserved sequences remains to be determined.

(c) Molecular characteristics of helianthinin and its precursors

The precursor polypeptide predicted from the *HaG3* sequence is 493 aa and has an M_r of 64.5 kDa. As with most legumin-like seed proteins, the *HaG3* gene product is rich in amide amino acids, e.g., glutamine and asparagine, and is relatively deficient in methionine and cysteine (Table I). As expected from previous 2D PAGE analyses (Allen, 1986), charged amino acids are distributed within the precursor polypeptide so that the α polypeptide has a net negative charge at neutral pH whereas the β polypeptide is positively charged under the same conditions.

The mechanism of post-translational processing and targeting of 11S globulins to protein bodies is complex, and although in some cases sequences required for these events are phylogenetically conserved (Borrito and Dure, 1987), the molecular basis of these events remains to be elucidated. The initial processing event, cleavage of the signal peptide, occurs co-translationally and results in the transport of the cleaved polypeptide into the lumen of the ER.

TABLE I

Amino acid composition of *HaG3* precursor polypeptide as predicted from the sequence in Fig. 2

Amino acid	Number (%)	Amino acid	Number (%)
Ala	38 (7.71)	Met	3 (0.60)
Cys	6 (1.22)	Asn	34 (6.90)
Asp	16 (3.25)	Pro	22 (4.46)
Glu	31 (6.29)	Gln	69 (14.0)
Phe	25 (5.07)	Arg	39 (7.91)
Gly	35 (7.10)	Ser	31 (6.29)
His	8 (1.62)	Thr	23 (4.66)
Ile	24 (4.87)	Val	29 (5.88)
Lys	10 (2.02)	Tyr	5 (1.01)
Leu	37 (7.51)	Trp	8 (1.62)

The probable NH_2 terminal leader sequence of the *HaG3* precursor is indicated in Fig. 2; this site was selected using the -1, -3 rules defined by von Heijne (1986) for signal sequence cleavage site selection. The location of the predicted α/β cleavage site is boxed in Fig. 2 (see below).

(d) Divergent subfamilies encode helianthinin

Hybridization and restriction analysis of two nearly full-length cDNA recombinants, *Ha2* and *Ha10*, suggested that sunflower 11S seed proteins were encoded by two divergent subfamilies (Allen, 1986; Allen et al., 1987b). These subfamilies are designated *Ha2* and *Ha10*, corresponding to the cDNAs that distinguish each subfamily. Genomic blot analyses (Allen, 1986; Allen et al., 1987b) revealed that the *Ha2* subfamily includes at least three members and the *Ha10* subfamily includes two or more members. Genomic sequences representative of each subfamily were isolated from a sunflower genomic DNA library; restriction maps of these recombinants are shown in Fig. 1. Regions that are complementary to either *Ha2* or *Ha10* are indicated. Even at relaxed hybridization criteria ($6 \times \text{SET}$, 55°C), *Ha2* and *Ha10*, or their genomic homologues, cross-hybridize very poorly (data not shown). Based on the intrafamilial sequence variation reflected in restriction site locations in regions flanking helianthinin coding sequences (Fig. 1), we conclude that *HaG10-B1* and *C1* are non-allelic members of the *Ha10* subfamily; similarly, *HaG3-A1* and *D1* are non-allelic members of the *Ha2* subfamily. Based on genomic blot analysis (Allen, 1986; Allen et al., 1987b), the helianthinin genes shown in Fig. 1 cannot represent all members of each subfamily. In the *Ha2* subfamily, at least two additional members remain uncharacterized, and in the *Ha10* subfamily, there is at least one additional member.

The extent of divergence between the *Ha2* and *Ha10* subfamilies is illustrated in Fig. 3 where the DNA sequence from a region of *HaG3*, including the α/β cleavage site (Fig. 2), is compared to a similar region of the cDNA, *Ha10*. Overall these nucleotide sequences share only 50% sequence similarity. The predicted *Ha10* and *HaG3* aa sequences share 43% similarity (data not shown). This latter observation suggests that the majority of the helianthinin coding sequence has diverged significantly, so much so that

```

1 .....30GTTTCAGGACAAAGGAAAGACAAGGCAACACAGATTTCAC 49
1480 TTCAGAGATTCTTCTTGGGCGGAAACGCTCAAGCTCAAGCTCAAGGCGAG 1529
46 GGACAACAAAGCAGACAACAGGAAGACAACAGGAGGAGAGAGTGC 95
1530 CAGCAACAAAGCAGACAACAGGCGGAAATCTCTCAAGGCGAAGGCA 1579
96 CTTTCCGCGGAGGAGGAA...CTGCACAGACACAATGTATAGGCTGGT 142
1580 AAGGCAAAAGGCAAGGCAAGGTCAGAACGCGGCAACATCTTCAAGGTT 1629
143 TCGATACTGAATTACTGCAGAGGGGTTTAAACGAGTGGAGGCTCAATC 192
1630 TCACGCGCGAGGCTGATTGCAATCATTTCAAGTGGACCAAGAGACCGGC 1579
193 ATCAGGGGAGTGGAGGAGTGCAGAACCGCGGTTTATGTCAGGTAGA 242
1640 CACAAGCTACAGGAGCAAAAGCAGCAGAGAGCGCCAGATTGTTAATGTCG 1729
243 GCAACAGATGGAAATCGTCACCGCTGAGGAGCA...AC 277
1730 ACAAGAGCTTCAATTAAGTCGCGGCGACCAAGACAGAGCGTCTCTCGGC 1779
278 AACACAAATGC...ACCAAGTAGGATCAAGCA 308
1780 AACACAAAGAGGAGGAGGAGGCTCTCTCGGCAACAAAGAGCAGAGCA 1829
309 GGAGGAGCCATC...CAAGGCTGTGGAAGAAACATATGAGTGC 350
1830 GCGAGAGCTGGGCGATGGAGCAAGGCTGTGGAAGAAACATATGAGGAT 1879
      N O V E R T I C S
351 TAAAGCTTCTGTACAACTTUGATAACCAAGAGAGGCTGATGTTCA 400
1880 GAAGTTCAAAGTGAACA...TTCAGAGCGTTCGAGGCTGACTTGTAA 1926
401 ACCGCGAAGCTGGAAACTCAACATGCTCAAGCAACAACTGCCAT 450
1927 ACCGCGAAGCGCGGAGGATTGCAAGCTCAACAGGTTCAAAATGCCAT 1976
451 GGTATCTCTGATGAGGCTCAATCGCGAGAAAGGAGAGCTCAACGGAATC 500
1977 CTGAGGAGCTCTCGGCTCAGGCTGGAAGAGGCGAACTCGGTGCGAATC 2025
501 CATTATTCTCGGACACTGCAAGTCAACAGGCAACAGCTGCTACAGT 350
2026 CGATCGAATCGGACACTGCAAGTCAACAGGCGAACTCTGCTAG. GT 2074
551 GGTCAACGGAGAGCGACAGATCGAAGTGGTGTGGAACAGGTTGAAGTC 600
2075 AACCGAGGAGGCTTGAAGCTCAAAATGCTGACAGCAAGCAAACTCAG 2124
601 TGTGAAAGGAGCAAGTCAAGAGGTTGACATTTTCCAGTGGCAGATTC 650
2125 TTTTGCACAAAGAGCTCGCTGAGGAGAGGTTGCTGCTGATCCGACAAAC 2174
651 CTTGCTCAGCAACTGCTCGAGTGGACAGAAATCGGTTCCAGTGGGTGCG 700
2175 TTTGCG...GTGATCAAGAGAGCGAATGAACAGGAGGAGGCTGCT 2221
701 GTTCAAGCAACAGGCTGCACTGAAGAGCGCA...TTAGCGCGGTACAGAT 749
2222 TTTCAAGCAATATCATATGCAATGATAGCAAACTTGCAGCGCGTGTGT 2271
750 CGGTTTTCGAGGCGATGCGGTTGAGGCTGATCAGCAACTGCTATCAGGT 799
2272 CCGCATCAGCAGCATCGGCTTGAAGTTGTGCGGCAATCGGTATCAGGTA 2321
800 TCACCGAACCAGGCTCAGAGCTTGAAGTCAACAGGAGAGCGAGCGCT 849
2322 TCTGACAGCAAGGCTCAGCAGCTCAAGTT...TAGCCAGAGGAGACCGT 2368
850 ACTGTTTTCTCCAGAGAGGAGTACTAGGCGAAGTAAATGTCGAGTAC 899
2369 TTGTGTCAGCA...AGTTTTCAGGCGGCA... 2399

```

Fig. 3. Comparison of *HaG3* and *Ha10* sequences. Nucleotide sequences of *Ha10* (upper sequence) and *HaG3* (lower sequence) spanning the region encoding the α/β cleavage site were compared. The aa sequence of the α/β cleavage site (boxed in Fig. 2) is shown below the nucleotide sequence. Solid bars indicate identical nucleotides. Dots indicate gaps inserted to maximize sequence homology. The first 1479 nt of the *HaG3* sequence are not shown.

second-hit mutations contribute significantly to the overall divergence of the two helianthinin sub-families. Although highly divergent throughout most of the protein coding sequence, the similarity of the DNA sequences encoding the α/β cleavage site approaches 90%; in this region, 24 of 27 nt are identical. The three nt differences in Fig. 3 are third base changes; consequently, the predicted α/β cleavage sites of *HaG3* and *Ha10* are identical.

(e) The α/β cleavage site is phylogenetically conserved

Comparison of the predicted aa sequences for sunflower helianthinin and *Brassica cruciferin* (Simon et al., 1985) revealed an overall similarity of 46% including conservative aa differences (data not shown). However, the region encoding the α/β cleavage site in helianthinin and cruciferin are nearly identical, differing by a conservative change from valine to leucine. The phylogenetic conservation at the α/β cleavage site is illustrated in Fig. 4. The α/β cleavage site and the nt sequence encoding each site for *HaG3* and *Ha10* and *Ha2* (Allen, 1986) are included. The α/β cleavage sequences for legumin-like seed proteins in seven other species, including both monocots and dicots, are also summarized in Fig. 4. At the aa level, the sequence conservation is striking. The presence of a serine in the last position appears to be characteristic of *legB*-type genes (Wobus et al., 1986); threonine at this position is indicative of *legA*-type genes. Based on these data and the intron/exon structure, we conclude that the *HaG3* helianthinin gene is a B-type gene and is most similar to the *LeB4* gene of *V. faba* (Baumlein et al., 1986) and *legJ/K* genes of pea (Gatchouse et al., 1988) than to the prototypical *legA* gene of pea (Lycett et al., 1984). Thus far, all helianthinin genes or cDNAs analyzed are of the *legB*-type.

(f) Conclusions

(1) The sunflower helianthinins are legumin-like seed proteins and are encoded by at least two divergent gene families defined by the sequences of *HaG3* and *Ha10*.

(2) Among legumin-like seed proteins of diverse plant species, the most conserved aa sequences are those required for appropriate post-translational

<u>Helianthus annuus</u>	↓ N G V E E T I C S
HaG3	AACGGTGTGGAAGAAACCATCTGCAGC
Ha2	AACGGTGTGGAAGAAACCATtTGCAGt
Ha10	AACGGTGTGGAAGAAACaATaTGCAGt
<u>Vicia faba</u>	
leg A	L V T AAtGGGcTtGAgGAAACCGTtTGCAct
leg B	L AAtGGTtTGGAAGAAACCATCTGtAGt
<u>Pisum sativum</u>	
leg A	L V T AAtGGGcTtGAgGAAACagTtTGCAct
leg J/K	L AAtGGTtTGGAAGAAACtATCTGtAGt
<u>Brassica napus</u>	L AACGGTtTaGAAGAgACCATaTGCAGC
<u>Arabidopsis thaliana</u>	
CRA-1	L AAtGGcTtTaGAgGAgACCATCTGCAGC
CRB	L L T AAtGGTtTaGAgGAgACttTgTGCAcC
<u>Gossypium hirsutum</u>	
Subfamily A	L F AAtGGccTcGAgGAAACttTCTGCctcC
Subfamily B	L F AACGGcTtTaGAAGAAACatTCTGCtca
<u>Avena sativa</u>	L N F AAtGGTtTGGAgGAgAttTCTGtcca
<u>Oryza sativa</u>	L D F T AACGGTtTGGAtGAgACGtTtTGCAcC
CONSENSUS	N G L E E T I C S F T

Fig. 4. Phylogenetic comparison of legumin α/β cleavage sequences. The complete aa sequence for the sunflower α/β cleavage site is shown; downward arrow indicates cleavage site. Only aa that differ from the sunflower sequence are shown for the other plant species. The nucleotide sequences encoding the α/β cleavage sites are displayed immediately below its corresponding complete or partial aa sequence. Nucleotides that differ from the HaG3 sequence are shown in lower case letters. A consensus α/β cleavage site is indicated at the bottom of the figure. Data sources include *Helianthus annuus*: this work; Allen et al. (1987b); Allen (1986); *Vicia faba*: Wobus et al. (1986); *Pisum sativum*: Lycett et al. (1984), Gatehouse et al. (1988); *Brassica napus*: Simon et al. (1985); *Arabidopsis thaliana*: Pang et al. (1988); *Gossypium hirsutum*: Chlan et al. (1986); *Avena sativa*: Walburg et al. (1986), B. Larkins, pers. communic.; *Oryza sativa*: Takaiwa et al. (1987).

processing and intracellular trafficking. With these constraints, the bulk of the aa sequence of the storage proteins apparently are free to diverge. Additional *cis*-acting regulatory sequences must also be 'functionally conserved' to ensure appropriate transcriptional regulation of legumin-like genes.

(3) Based on the aa sequence of the α/β cleavage site, the seed protein encoded by *HaG3* is most similar to the *V. faba* *LeB4* gene (Baumlein et al., 1986) and the pea *legJ/K* genes (Gatehouse et al., 1988). Furthermore, based on the diversity of intron number and location among various legumin-like storage protein genes, it is likely that the progenitor gene for the B-type legumin genes was an A-type legumin gene (Lycett et al., 1984) containing three introns.

ACKNOWLEDGEMENTS

This research was supported by grants from the Texas Advanced Technology Research Program and Rhône-Poulenc Agrochimie. R.D.A. was a recipient of a W.R. Grace predoctoral fellowship. We thank Drs. Carl Adams and Juan Jordano for their critical review of this manuscript and also Concepcion Almoguera for critical technical assistance. We thank P. Pang, R. Pruitt and E. Meyerowitz for sharing their results prior to publication.

REFERENCES

- Allen, R.D.: Expression of 11S Seed Storage Protein Genes of *Helianthus annuus* L. Ph.D. Dissertation, Texas A&M University, 1986.
- Allen, R.D., Nessler, C.L. and Thomas, T.L.: Developmental expression of sunflower 11S storage protein genes. *Plant Mol. Biol.* 5 (1985) 165-173.
- Allen, R.D., Cohen, E.A., Vonder Haar, R.A., Adams, C.A., Ma, D.P., Nessler, C.L. and Thomas, T.L.: Sequence and expression of an albumin storage protein in sunflower. *Mol. Gen. Genet.* 210 (1987a) 211-218.
- Allen, R.D., Cohen, E.A., Vonder Haar, R.A., Orth, K.A., Ma, D.P., Nessler, C.L. and Thomas, T.L.: Expression of embryo specific genes in sunflower. In Davidson, E.H. and Firtel, R. (Eds.), *Molecular Approaches to Developmental Biology*, Alan R. Liss, New York, (1987b) pp. 415-424.
- Benton, W.D. and Davis, R.W.: Screening λ gt recombinant clones by hybridization to single plaques in situ. *Science* 196 (1977) 180-182.
- Baumlein, H., Wobus, U., Pustell, J. and Kafatos, F.: The legumin gene family: Structure of a B type gene of *Vicia faba* and a possible legumin gene specific regulatory element. *Nucleic Acids Res.* 14 (1986) 2707-2720.
- Borrotto, K. and Dure III, L.: The globulin seed storage proteins of flowering plants are derived from two ancestral genes. *Plant Mol. Biol.* 8 (1987) 113-131.
- Chian, C.A., Pyle, J.B., Legocki, A.B. and Dure III, L.: Developmental biochemistry of cottonseed embryogenesis and germination, XVIII. cDNA and amino acid sequences of members of the storage protein families. *Plant Mol. Biol.* 7 (1986) 475-489.
- Cohen, E.A.: Analysis of Sunflower 2S Seed Storage Protein Genes. M.S. Thesis, Texas A&M University, 1986.
- Dale, R.M.K., McClure, B.A. and Houchins, J.P.: A rapid single-stranded cloning strategy for producing a sequential series of overlapping clones for use in DNA sequencing: applications to sequencing the corn 18S rDNA. *Plasmid* 13 (1985) 31-40.
- Devereux, C., Haeblerli, P. and Smithies, O.: A comprehensive set of sequences and analysis programs for the VAX. *Nucleic Acids Res.* 12 (1984) 387-395.
- Favaloro, J., Treisman, R. and Kamen, R.: Transcription maps of polyoma virus specific RNA: Analysis by two dimensional nuclease S1 mapping. *Methods Enzymol.* 65 (1980) 718-749.
- Frishauf, A.M., Lehrach, H., Poustka, A. and Murray, N.: Lambda replacement vectors carrying polylinker sequences. *J. Mol. Biol.* 170 (1983) 827-842.
- Gatehouse, J.A., Bown, D., Gilroy, J., Levasseur, M., Castleton, J. and Ellis, T.H.N.: Two genes encoding 'minor' legumin polypeptides in pea (*Pisum sativum* L.). *Bioch. J.* 250 (1988) 15-24.
- Goldberg, R.B.: Regulation of plant gene expression. *Phil. Trans. R. Soc. B* 314 (1986) 343-353.
- Higgins, T.J.V.: Synthesis and regulation of major proteins in seeds. *Annu. Rev. Plant Physiol.* 35 (1984) 191-221.
- Lycett, G.W., Croy, R.R.D., Shirsat, A.G. and Boulter, D.: The complete nucleotide sequence of a legumin gene from pea (*Pisum sativum* L.). *Nucleic Acids Res.* 12 (1984) 4493-4506.
- Maniatis, T., Fritsch, E.F. and Sambrook, J.: *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1982.
- Maxam, A. and Gilbert, W.: Sequencing end-labeled DNA. *Methods Enzymol.* 65 (1980) 499-560.
- Messing, J.: New M13 vectors for cloning. *Methods Enzymol.* 101 (1983) 20-78.
- Mount, S.: A catalogue of splice junction sequences. *Nucleic Acids Res.* 10 (1982) 459-472.
- Pang, P.P., Pruitt, R.E. and Meyerowitz, E.M.: Molecular cloning, genomic organization, expression and evolution 12S seed storage protein genes of *Arabidopsis thaliana*. Submitted.
- Sanger, F., Nicklen, S. and Coulson, A.R.: DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74 (1977) 5463-5467.
- Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. and Roe, B.A.: Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *J. Mol. Biol.* 143 (1980) 161-178.
- Shotwell, M.A. and Larkins, B.A.: The biochemistry and molecular biology of seed storage proteins. In Marcus, A. (Ed.), *The Biochemistry of Plants: A Comprehensive Treatise*, Vol.

Pla
pro
Simon
anc
Br
leg
19
Takai
fan
tw
von H
cle

- ...nt Molecular Biology, Academic Press, New York, in
...ss.
- Simon, A.E., Tenbarger, K.M., Scofield, S.R., Finkelstein, R.R.
and Crouch, M.L.: Nucleotide sequence of a cDNA clone of
Brassica napus 12S storage protein shows homology with
legumin from *Pisum sativum*. *Plant Mol. Biol.* 5 (1985)
191-201.
- Takaiwa, F., Kikuchi, S. and Oono, K.: A rice glutelin gene
family: a major type of glutelin mRNA can be divided into
two classes. *Mol. Gen. Genet.* 208 (1987) 15-22.
- von Heijne, G.: A new method for predicting signal sequence
cleavage sites. *Nucleic Acids Res.* 14 (1986) 4683-4690.
- Walburg, G. and Larkins, B.A.: Isolation and characterization of
cDNAs encoding oat 12S globulin mRNAs. *Plant Mol. Biol.*
6 (1986) 161-169.
- Wobus, U., Baumlein, H., Bassünner, R., Heim, U., Jung, R.,
Müntz, K., Saalbach, G. and Weschke, W.: Characteristics
of two types of legumin genes in the field bean (*Vicia faba* L.
var. minor) genome as revealed by cDNA analysis. *FEBS
Lett.* 201 (1986) 74-80.

Communicated by T.D. McKnight.

The sequence of a gene encoding convicilin from pea (*Pisum sativum* L.) shows that convicilin differs from vicilin by an insertion near the *N*-terminus

David BOWN,* T. H. Noel ELLIS† and John A. GATEHOUSE*‡

*Department of Botany, University of Durham, South Road, Durham DH1 1LE, and †John Innes Institute, Colney Lane, Norwich NR4 7UH, U.K.

NOTICE: This Material
may be protected by copyright
law. (Title 17 U.S. Code)

The sequence of a gene encoding convicilin, a seed storage protein in pea (*Pisum sativum* L.), is reported. This gene, designated *cvcA*, is one of a sub-family of two active genes. The transcription start of *cvcA* was mapped. Convicilin genes are expressed in developing pea seed cotyledons, with maximum levels of the corresponding mRNA species present at 16-18 days after flowering. The gene sequence shows that convicilin is similar to vicilin, but differs by the insertion of a 121-amino-acid sequence near the *N*-terminus of the protein. This inserted sequence is very hydrophilic and has a high proportion of charged and acidic residues; it is of a similar amino acid composition to the sequences found near the C-terminal of the α -subunit in pea legumin genes, but is not directly homologous with them. Comparison of this sequence with the 'inserted' sequence in soya-bean (*Glycine max*) conglycinin (a homologous vicilin-type protein) suggests that the two insertions were independent events. The 5' flanking sequence of the gene contains several putative regulatory elements, besides a consensus promoter sequence.

INTRODUCTION

Convicilin has been termed a 'third storage protein' in pea seeds, in addition to legumin and vicilin [1]. It can be purified from both legumin and vicilin, and it consists solely of polypeptides of M_r approx. 71 000. It does not thus contain polypeptides found in either of the two major storage proteins [2]. On the other hand, convicilin is antigenically similar to vicilin [1], and it is possible to produce molecules containing both vicilin and convicilin polypeptides; for this reason, some authors have considered that convicilin and vicilin are the same protein [3]. Sequence data for a partial cDNA clone, pCD 59, identified as encoding convicilin by hybrid-release translation, supported this view, since the deduced amino acid sequence was strongly homologous with that of vicilin [4,5]. However, pCD 59 did not hybridize to vicilin cDNA species [5] or vicilin genes [6].

Variation in the mobility of convicilin polypeptides, on SDS/polyacrylamide-gel electrophoresis, between pea lines has allowed a convicilin locus, designated '*cvc*', to be mapped to chromosome 2 in pea [7]; it is distinct from any vicilin locus so far identified [8,9]. Convicilin has been shown to be encoded by a small gene family: hybridization of the cDNA clones pCD 59 and pCD 75 (a longer version of pCD 59; [5]) to genomic DNA restricted with endonucleases detected one or two hybridizing fragments, depending on which probe was used [5,6,9].

The isolation of a genomic clone containing a convicilin gene, putatively corresponding to the *cvc* locus, has been described [9]. The present paper reports the sequence of this gene and its flanking regions, and shows that convicilin genes in pea (*Pisum sativum* L.) form a sub-family of the total family of vicilin-type genes.

MATERIALS AND METHODS

Materials

Pea seeds of the cultivar (cv.) Feltham First were obtained from Suttons Seeds, Torquay, Devon, U.K.; seeds of cv. Dark Skinned Perfection were from S. Dobie and Son, Torquay, Devon, U.K. The isolation of the genomic clone lambda JC4, and its sub-clone pJC 4-100, from a genomic library prepared from DNA isolated from *Pisum sativum* cv. Dark Skinned Perfection has been described previously [9]. Reagents and enzymes for M13 DNA sequencing were from Gibco/BRL (Gibco, Paisley, Renfrewshire, Scotland, U.K.); restriction enzymes were supplied by Northumbrian Biologicals, Cramlington, Northd., U.K. S1 nuclease and other enzymes were from BCL, Lewes, East Sussex, U.K. Radiochemicals were supplied by Amersham International, Amersham, Bucks., U.K. Other reagents used were of analytical quality wherever possible. Nitrocellulose filters were type BA85 (Schleicher und Schuell) from Anderman and Co., East Molesey, Surrey, U.K.

Methods

DNA sequencing. Restriction mapping on pJC 4-100 was carried out by conventional methods [10]. Preparation of subclones from pJC 4-100 in pUC18 or pUC19, preparation of sequencing subclones in M13 mp18 or mp19, preparation of single-stranded DNA, and dideoxynucleotide DNA sequencing using [α -³²S]thio-dATP were also carried out by standard techniques [11-14]. The sequence given was determined by overlapping sequences from subclones; both strands of the DNA were fully sequenced. Sequences were analysed by diagonal dot-matrix comparisons [15], using a

These sequence data have been submitted to the EMBL/GenBank Data Libraries under the accession number Y00721.
‡ To whom correspondence and reprint requests should be addressed.

program written by ourselves and by manual comparisons supplemented by sequence-handling software (programs NNALN and FASTP, kindly supplied by Dr. W. Pearson). Hydrophobicity profiles were plotted using the method of Hopp & Wood [16].

Blotting techniques. Restriction fragments from pJC 4-100 or its subclones were isolated from low-gelling-temperature agarose gels [17] and labelled with [α - 32 P]dCTP (400 Ci/mmol; 100 μ Ci used/0.2–0.5 μ g of DNA) by nick translation [18]. 'Southern' blots of agarose-gel separations of restriction fragments, or digests of pea leaf genomic DNA (purified as in [19]) with restriction enzymes, were prepared and hybridized to denatured labelled probes in $5\times$ SSC (1 \times SSC is 0.15 M NaCl/0.015 M sodium citrate buffer, pH 7.2)/2 \times Denhardt's solution (1 \times Denhardt's solution is 0.02% Ficoll/0.02% bovine serum albumin/0.02% polyvinylpyrrolidone)/denatured herring sperm DNA (100 μ g/ml), at 65 $^{\circ}$ C as described in [20]; subsequent washes were to a hybridization stringency of $0.1\times$ SSC at 65 $^{\circ}$ C. 'Northern' blots of agarose-gel separations of glyoxalated total RNA samples (prepared from pea (cv. Feltham First) cotyledons at different developmental stages as previously described [21]) were prepared and hybridized to denatured labelled probes in $5\times$ SSC, 2 \times Denhardt's solution/denatured herring sperm DNA (200 μ g/ml/50% (v/v) formamide, at 42 $^{\circ}$ C [22]; subsequent washes were to a hybridization stringency of $0.1\times$ SSC/0.1% SDS at 50 $^{\circ}$ C. Densitometry of autoradiographs, obtained by exposing the washed blots to preflashed X-ray film at -80° C, was carried out on an I.K.B. (Bromma, Sweden) Ultrosan XL densitometer.

S1 mapping. S1 mapping was carried out as described by Favaloro *et al.* [23]. Each assay mixture contained 5 μ g of polyadenylated RNA, prepared from pea (cv. Feltham First) cotyledons at a mid-development stage (14–15 days after flowering) as previously described [24], and at least 0.2 μ g (approx. 2×10^6 c.p.m.) of DNA probe, 5' end-labelled [25] with [γ - 32 P]-ATP (6000 Ci/mmol; 50 μ Ci used/0.2–0.5 μ g of DNA). The protected fragment after S1 digestion was run on a DNA sequencing gel, and its 3' end was mapped by running a DNA sequencing reaction that covered the same region of sequence on the same strand, and had been primed by an oligonucleotide primer whose 5' end corresponded to the site of labeling, in adjacent tracks. Controls omitting RNA were carried out.

Protein sequencing. Convicillin was purified as previously described [1]. Portions (2 mg) of the protein, dissolved in 0.1% trifluoroacetic acid, were subjected to h.p.l.c. (Vydac reverse-phase C_{18} column; elution with a gradient of acetonitrile in 0.1% trifluoroacetic acid) to remove traces of vicilin. Convicillin polypeptides were digested with trypsin, and the resulting peptides were separated by h.p.l.c. and sequenced by the manual diaminobenzoyl isothiocyanate method, as previously described [26]. N-Terminal sequences for convicillin were obtained by automated sequence determination on an Applied Biosystems model 371A protein sequencer, with online h.p.l.c. residue identification. A 0.3 mg sample of protein was used per determination.

RESULTS

Genomic clone

A partial restriction map for the genomic subclone pJC 4-100 has been published previously [9]. A revised and detailed map, showing the position of the gene and the region sequenced, is given in Fig. 1(a). The clone contains approx. 8 kb of sequence 5' flanking the convicillin coding sequence, and approx. 3 kb of 3' flanking sequence; these regions do not contain sequences hybridizing to probes from the *cvcA* coding sequence (results not shown). Regions of this clone outside the sequenced region are not discussed further in the present paper.

The convicillin gene

The sequencing map for the convicillin gene is given in Fig. 1(b), and the complete sequence of the gene and its immediate 3' and 5' flanking regions is given in Fig. 2. We have designated this gene '*cvcA*'. The predicted sequence of the encoded protein was deduced by homology with vicilin and by the presence of an open reading frame at the 5' end, and is also shown in Fig. 2. The coding nucleotide sequence is interrupted by five introns, whose positions could be inferred from the predicted and determined protein sequence (the present paper) and from the nucleotide sequences of the convicillin cDNA species pCD 59 [5], the homologous *Phaseolus vulgaris* (French bean) vicilin (phaseolin) gene [27] and homologous pea vicilin cDNA species and genes ([28,29], J. A. Gatchouse, D. Bown, M. Levasseur, R. Sawyer & T. H. N. Ellis, unpublished work). The sequence from start codon to stop codon thus contains six exons, of 661, 176, 75, 324, 283 and 197 bases respectively, and five introns, of 151, 103, 103, 88 and 97 bases respectively. The encoded amino acid sequence is 571 amino acids in length, and predicts a precursor polypeptide of M_r 66986; when the leader sequence of 28 amino acids (see below) is subtracted the predicted M_r for the mature polypeptide is 63928. The discrepancy between this value and the polypeptide M_r determined for convicillin (71000) is discussed below.

The 3' flanking sequence given extends for 428 bases after the stop codon; a further 450 bases of sequence have been determined, but do not show any significant features and will not be discussed further. Two polyadenylation sites are present in the 3' flanking sequence: 119 and 134 bases after the stop codon; the first of these is of the multiple overlapping type (AATAAATAAA) often found in plant genes [30]. The 5' flanking sequence contains a good match to the consensus sequence for a plant gene 'TATA' box [31] 66 bases before the start codon (CTATAAATA). Other sequence features in this region are discussed below.

Partial sequence of convicillin

The identity of the gene *cvcA* was confirmed by comparing its predicted protein sequence with partial protein sequence data from convicillin. In all, 16 residues at the N-terminus of convicillin and an additional 75 residues from 14 tryptic peptides were determined. Results are shown in Fig. 2. The determined sequences agree fully with the sequence predicted by *cvcA* and show that the first 28 residues of the predicted sequence are not present in the mature polypeptide. These removed residues constitute a typical 'leader' sequence [32]. At

Fig. 1. Re-

Key:
N, Nucleo-

amino ac-
peptides
notation
showing
valid.

Expressi-

An S1
the expres-
start. T
bases
labelled
polyaden-
cotyledon
nuclease
shown in
were ob-
sequence
region 2
designat-
in the S
the pro-
Protein
consens-
the abor-
bands

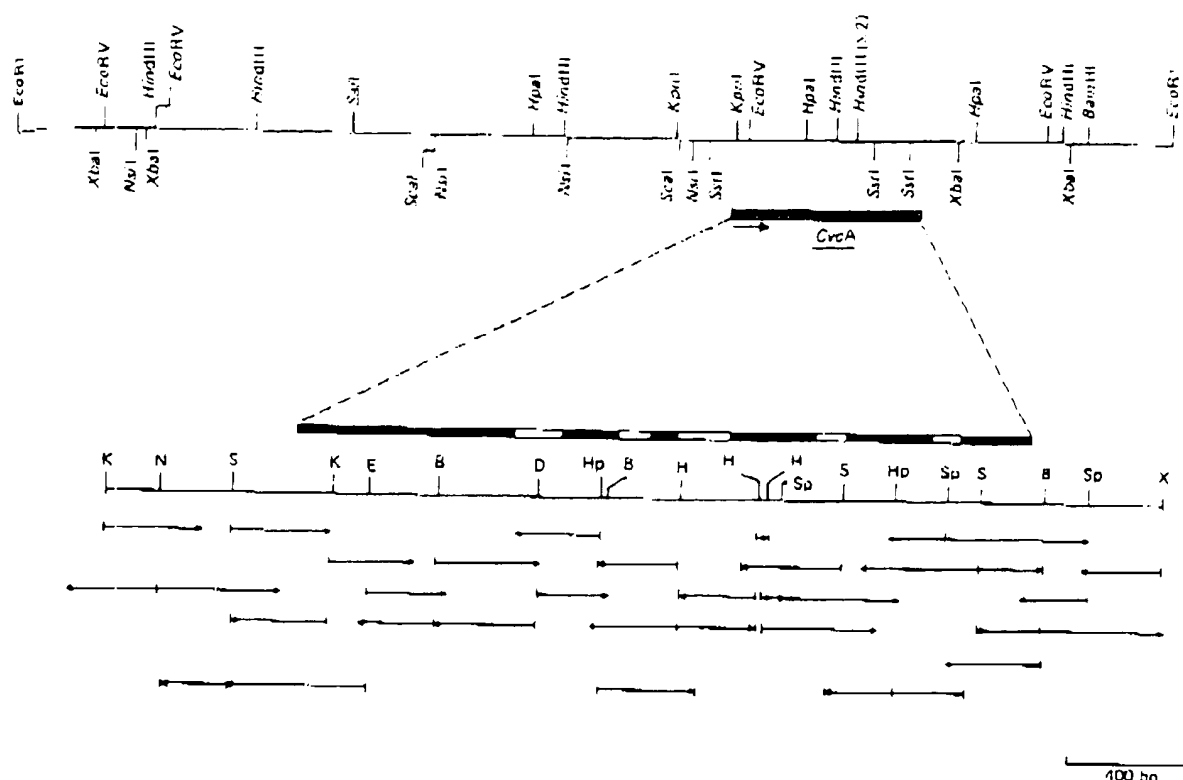


Fig. 1. Restriction map of the clone pJC4-100 containing *cvcA*, and sequencing map of *cvcA*

Key to restriction site symbols on sequencing map: B, *Bgl*II; D, *Dra*I; E, *Eco*RV; H, *Hind*III; Hp, *Hpa*I; K, *Kpn*I (= *Asp*7181; N, *Nsi*I; S, *Sma*I; Sp, *Ssp*I; X, *Xba*I.

amino acid 209, two residues were found in tryptic peptides; N, as predicted by *cvcA*, and Q (one-letter notation). Peptides were obtained from all six exons, showing that the assignment of intron positions was valid.

Expression of *cvcA*

An S1 mapping experiment was carried out to confirm the expression of *cvcA* and to locate the transcription start. The *Asp*7181 restriction fragment, covering bases -561 to 143 in *cvcA*, was isolated and 5'-end-labelled. After hybridization of the labelled fragment to polyadenylated RNA isolated from developing pea cotyledons, the nucleic acids were treated with S1 nuclease and analysed by gel electrophoresis. Results are shown in Fig. 3(a). Protected fragments of 139-150 bases were obtained, suggesting that an mRNA had identical sequence with the probe from base 143 in *cvcA* to a region 24-35 bases 5' to the ATG start codon. The base designated '+1' was that giving the most intense band in the S1 mapping assay, i.e. the underlined base in the protected sequence region, CATCATCTAAAG. Protected fragments extending to the A bases in the consensus transcription start sequences -CATC- [31] in the above region were observed, but gave less intense bands in the S1 mapping assay. Control experiments

with no RNA present gave no protected fragment. A further S1 mapping experiment, with the *Nsi*I-*Eco*RV restriction fragment, covering bases -382 to 257 in *cvcA*, gave protected fragments ending in the region -8 to +2. In this case both the S1 mapping assay and its control with no RNA present gave protected fragments corresponding in length to the original probe.

The developmental expression of convicilin genes was also studied by hybridization of part of the sequence of this gene to total RNA prepared from pea cotyledons at different stages of seed development. The probe fragment was chosen to include only the 5'-end of the coding sequence of the gene to avoid cross-hybridization to vicilin mRNA species. Pea cotyledon RNA was glyoxalated, size-fractionated by electrophoresis and blotted on to nitrocellulose before hybridization to the *Sma*I-*Bgl*II (bases -176 to 462) fragment of *cvcA*, labelled by nick translation. The results of this experiment are shown in Fig. 3(b). The probe hybridized to two bands of similar mobility on the Northern blot, corresponding to mRNA species of approx. 2650 and 2500 bases; the larger of the two species consistently gave a more intense hybridization signal, the ratio of the integrated peak areas of the two bands being approx. 3:1 (± 0.7) in all tracks. No evidence of hybridization to vicilin mRNA species, which have been previously

Sequence of <i>pca</i> convicilin gene	
AAAGTAC -507	CycA TCAGGGAACAAATGAGGAATTGAGAAAGCTTGCAAAATCAAGCTCAAGGAAAGAACTTACCCCTGGAATTTGAACCTTTCACTTGAGAAAGCCACAAGCCAGAAATATTCTAATAAGTT 1534 A.A. S H E D I E E L R K L A K S S S K K S L P S E F E P F M L H S H K P E Y S M K F
TAATTA -387	CycA GGCAGGTTGTTTGAATTACTCCAGAGAAAAATACCTCAGCTTCAAGATTAGATATACITGTTAGTTGTGTGAGATTAAAGGATGTACACAACCTAAATATATATAAAGAC 1654 A.A. G K L F E I T P E K K Y P O L O D I L V S C V E I N K
CTTAAA -267	CycA CATTTTAATTATATTACAGAAATATGTTAATGCGTTTTGCTTAAATTTTAGGGAGCTCTAATGTTGCCACACTACAATTCAGGGCAATAGTTGTACTATTAGTTAAISAGGAAA 1774 A.A. IVS-4 G A L M L P H Y N S R A I V V L L V N E G K
CAATAC -147	CycA AGGAACCTTGAACTTCTGGGTTTAAAAATGAGCAACAAGAGAGGGAAGATAGAAAGAAAGAAACAATGAAGTGCAGAGATATGAAGTATGATTCGGGGTGAGCTGTTATCAT 1894 A.A. G H L E L L G L K N E Q Q E R E D R K E R N N E V Q R Y E A R L S P G D V V I I
ATATCT -27 20X...	CycA TCAGAGGTCACCCAGTTGCCATTAGTGTCTCATCAATCTGAATTTGCTTGGATTGGTATCAATGCCAAGAACATCAGAGAACTTCCTTTCAGGTATTAGTGAATAGTAATATC 2014 A.A. P R G H P V A I S A S S N L N L L G F G I N A K N N Q R N F L S
GGCTCT 94 A S	CycA ATTAGTTAATAATTTTCGATTAAATGAGAAATATTTGAATGTTATTTCTAATTTGGGGATTGAAATTTGAAGGATCGATGACAAATGTGATAAGCCAAATAGAAAATCCAGTAAGG 2134 A.A. IVS-5 G S D D N V I S Q I E N P V K
ACCTTCA 214 P S	CycA AGCTCAATTTCTCTGATCTTCTCAAGAGGTAAATAGATTAAATCAAGAAATCAAAACAATCTCACTTTGCAAGTGTGAACCTGAACAAAGGAGGAGAAAGCAAGAAAAAGGAGTC 2254 A.A. E L I F P G S S Q E V N R L I K N Q K Q S H F A S A E P E Q K E E S Q R K R S
GGAGCA 334 E U	CycA CTCTGCTTCAGTTCTGGACAGTTTTACTGAGTAATCAATATGAAAAATATGAGATGTATGAGCTAAGATCTAGCTAGCTCTTCTGAGCTAAGAGTAATAATGATCTTGTAACT 2374 A.A. P L S S V L D S F Y
GGAGAA 454 E E	CycA CTACCTATTGAGCCCACTTTTCTATACGAATAAATAAATAATTAATAAACTTGCTTTTTTTTACTTTAACTACAAGGATATTAATTTGTTGTTCTGGGGTAAGCTTAAAA 2494 A.A. (PolyA+) ... (PolyA+)
CAACGT 574 K R	CycA AAAGACTATGGATTCAATGAAGGAATTTTAAATTTTAAATATGTTTATGTTGTTTATTGTAATGTTTCAACATGACAGTCCCTACTCTTGTATTAGTTGCTTTAAT 2614
TTGTAAT 694 C	CycA TTGCTTTAATTTGTTTATGTTTTATATCTTTTCTTAAATTAATAAATGGAAGTGTGTTGTAATTTGTAAGTAAAGAGAGTTGCAATTTCTTTCTCTAGA 2723

Fig. 2. Sequence of gene *cycA* ('CycA'), with the predicted sequence of the convicilin precursor polypeptide ('A.A.')

The predicted site of cleavage of the leader sequence is indicated by a colon (:). The base designated +1 is indicated by a circumflex (^). Other sequence features are as indicated on the Figure. The N-terminal sequence determined for convicilin, and the sequences of convicilin tryptic peptides, are indicated by double and single underlinings respectively; vertical lines indicate the termini of the peptides.

Identified as approx. 1700 bases in size [33], was obtained, showing that the probe was specific for convicilin mRNA species. The relative intensities of the hybridizing bands from different developmental stages show that the proportion of convicilin mRNA species in total RNA increases as cotyledon expansion proceeds, to a maximum at 16–18 days after flowering, and decreases thereafter. The peak in convicilin mRNA levels agrees with previous observations that convicilin synthesis is maximal during the second half of cotyledon expansion [34].

Hybridization to genomic DNA

Pea genomic DNA from cvs. Feltham First and Dark Skinned Perfection was digested with various restriction enzymes, size-fractionated by agarose-gel electrophoresis

and blotted on to nitrocellulose. The blots were then hybridized with the labelled convicilin specific probe (*Sst*I–*Bgl*II; bases –176 to 462) described above. Results are shown in Fig. 4. The two cultivars gave identical band patterns in all restriction digests made. Digests with *Eco*RI gave two bands, one of approx. 13 kb, corresponding to the *Eco*RI fragment in pJC 4-100, and one of approx. 9.0 kb, corresponding to the *Eco*RI fragment previously identified as hybridizing to the convicilin cDNA species pCD 59 and pCD 75 [5]. Both these bands were present at an indicated level of approx. one copy per haploid genome, as shown by a reconstruction assay where gene copy equivalents of pJC 4-100 were hybridized on the same filters. All other restriction digests gave two or more hybridizing bands, consistent with the restriction

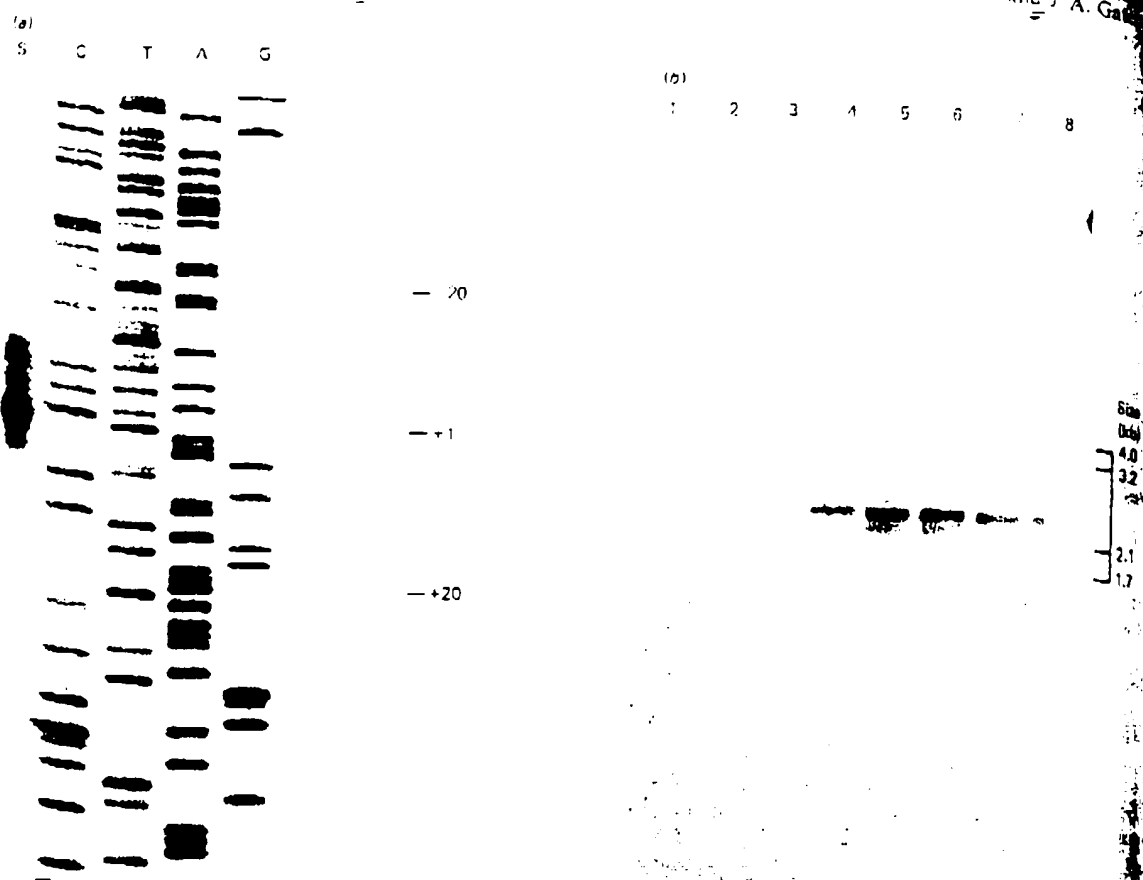


Fig. 3. Expression of convicilin gene *cvcA*

(a) S1 mapping experiment to locate the transcription start in *cvcA*. The protected fragment is run in track S; other tracks are the corresponding region of DNA sequence (the sequence is given in complement, and must be read down the sequencing gel). (b) 'Northern' blot, showing hybridization of *Sst*I-*Bgl*II probe (bases -176 to 462) from *cvcA* to total RNA isolated from developing pea cotyledons (line Feltham First) at 8 days after flowering (d.a.f.) (track 1), 10 d.a.f. (track 2), 12 d.a.f. (track 3), 14 d.a.f. (track 4), 16 d.a.f. (track 5), 18 d.a.f. (track 6), 20 d.a.f. (track 7) and 22 d.a.f. (track 8). Under these conditions the cotyledon expansion phase of development lasts from 7-8 d.a.f. to 21-22 d.a.f. [24,32]. A 10 μ g portion of total RNA was loaded per track in the original gel electrophoresis. The molecular-size scale is taken from standard RNA species (ribosomal RNAs) run on the original gel.

map of *cvcA* (see Fig. 1), at intensities consistent with the conclusion that two convicilin genes were present per haploid genome, in agreement with previous reports [6].

DISCUSSION

Coding sequence

The amino acid sequences predicted by *cvcA*, and found for convicilin, confirm the presence of a 'leader' sequence on the precursor polypeptide, as had been previously suggested by translation experiments *in vitro* [35]. The sequence for the mature polypeptide predicted by *cvcA* is then in good agreement with the amino acid composition of convicilin, as shown in Table 1. The presence of one methionine residue in the mature polypeptide is correctly predicted by *cvcA*, and its position (amino acid 388) is consistent with the observed results of CNBr cleavage of convicilin, which generates two fragments of approx. 55000 and 15000 *M*. [1].

Despite the evidence that *cvcA* is a convicilin gene and that it is expressed, it differs in its sequence from the convicilin cDNA identified by Domoney & Casey [4], which was used to select the genomic clone containing *cvcA*. The overall homology between the two sequences is 94% over 590 corresponding bases. The main difference between the two sequences is a deletion of 18 nucleotides (six amino acids) in pCD59 relative to *cvcA*, corresponding to a region near the hypothetical α -subunit processing site in vicilin [26]. There are also a number of conservative amino acid substitutions in the remainder of the sequence (not shown). These sequence differences are sufficient to account for the previous observation [5] that pCD 59 hybridized to only one of the two convicilin genes detected by the *cvcA* probe in the present study. The data suggest that pCD 59 represents the second convicilin gene detected by hybridization to genomic DNA, *cvcB*, which is thus shown to be functional. When pCD 59 was hybridized to RNA from developing pea cotyledons [5], only one band was detected

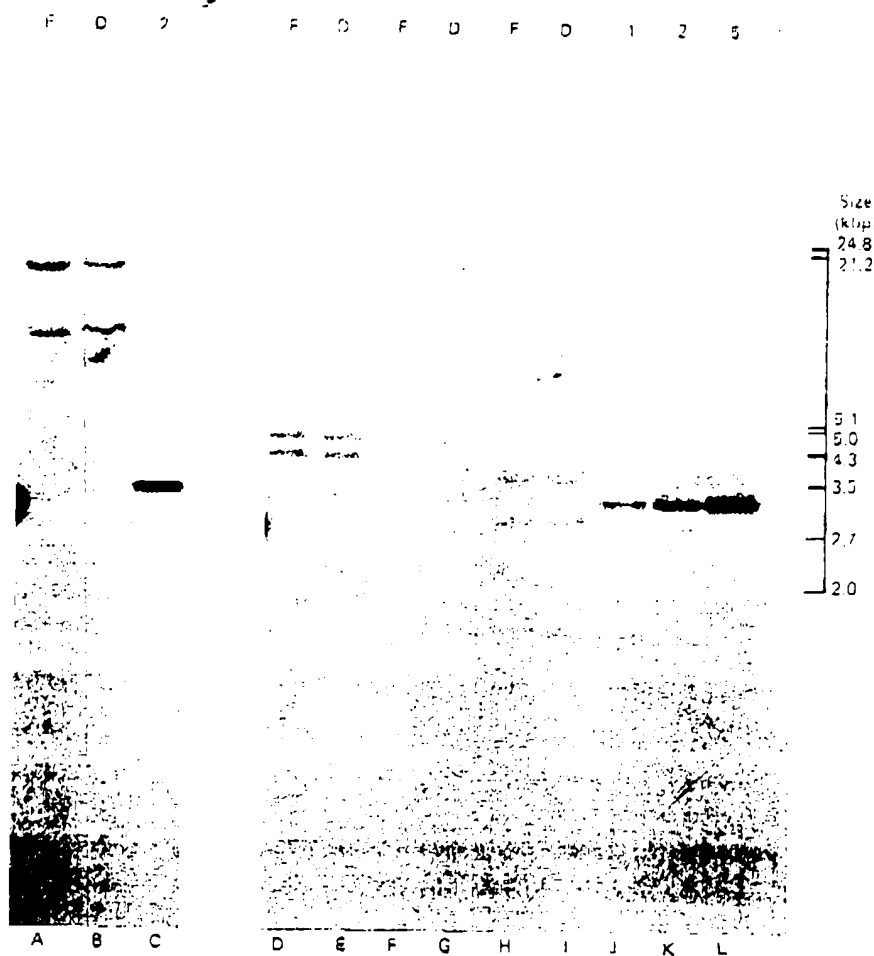


Fig. 4. Southern blot showing hybridization of *SsrII*-*BglII* probe (bases -176 to 462) from *cvcA* to restriction digests of genomic DNA from lines Feltham First (F) and Dark Skinned Perfection (D)

A 10 μ g portion of DNA was loaded per track on the original gel electrophoresis. Restriction enzymes used were as follows: A and B, *EcoRI*; D and E, *BglII*; F and G, *BamHI*; H and I, *EcoRV*. The blot is calibrated with gene equivalent amounts [33] of digested pJC4-100; the indicated copy numbers per haploid genome are given above tracks C, J, K and L. Tracks A-C are from a different gel to the remainder. The molecular-size scale is from restriction digests of standard DNA species run on the original gels.

in gene and
e from the
Casey [4],
containing
sequences
The main
action of 18
ive to *cvcA*,
hetical $\alpha:\beta$
are also a
tions in the
se sequence
ne previous
one of the
robe in the
represents
lization to
wn to be
RNA from
as detected

on a 'Northern' blot, as opposed to the two detected by the *cvcA* probe, suggesting that *cvcA* and *cvcB* each gives rise to a distinct mRNA species. Further data will be necessary to confirm this conclusion.

Homology with vicilin. A dot-matrix comparison of the polypeptide sequences predicted for convicilin, and for a vicilin 50000-M_r polypeptide is given in Fig. 5. The sequences are strongly homologous over most of their length, with short areas of low homology apparent at regions corresponding to the sequences around the putative $\alpha:\beta$ and $\beta:\gamma$ subunit processing sites in vicilin. These areas have previously been noted as being of low homology when pea vicilin polypeptides are compared with those from different species [28]. The major difference between the two sequences is apparent as a large insertion in the convicilin sequence near its N-terminus, corresponding to sequence being inserted between amino acids 3 and 6 of the mature vicilin polypeptide. Homology over the region -3 to +3 is

weak at the amino acid level, but significant at the nucleotide level; outside this region, and the insertion, homology is strong in both directions (see Fig. 5). The convicilin leader sequence is homologous with that in vicilin, but not to leader sequences in other seed proteins (results not shown), showing that the extra sequence in convicilin represents an insertion into a vicilin gene rather than a 5' addition to it. The strong homology of convicilin with vicilin outside the inserted sequence accounts for the overall similarity in properties between the two proteins and their antigenic similarity [1]; it would also account for their ability to form molecules containing polypeptides of both vicilin and convicilin.

The homology in amino acid and corresponding nucleotide sequences between *cvcA* and vicilin genes in pea (results not shown; homology at the nucleotide level between the vicilin cDNA pAD2.1 [29] and corresponding sequence regions in *cvcA* is 79%) shows that the *cvcA* gene should be regarded as belonging to a sub-family of the vicilin gene family; this designation supports both

Table 1. Amino acid composition of convicillin; comparison of predicted and experimental compositions

Amino acid	Residues predicted	Composition (mol/100 mol)	
		Predicted	Found*
D	23	59	10.87
N	36		
T	13	113	2.39
S	40		
E	80	20.81	2.55
Q	33		
P	25	5.90	6.39
G	27		
A	18	4.60	5.47
C	1		
V	27	4.23	5.90
M	1		
I	24	0.17	0.13
L	49		
Y	15	4.97	4.46
F	20		
W	3	0.13	4.46
K	43		
H	12	4.42	3.85
R	53		
		9.76	8.71
			2.59
			3.30
			ND†
			8.18
			2.22
			8.15

* From [1].

† ND, not determined.

previous views that convicillin was distinct from [1], was essentially the same as [3], vicilin.

Nature of the inserted sequence in convicillin. The inserted sequence in convicillin will be considered as amino acids (+)4-124 or nucleotides 121-483. At the amino acid level, the sequence contains a high proportion of charged and hydrophilic residues (from 121 amino acids, there are 38 glutamate residues, 24 arginine residues and 9 lysine residues; only 10 residues are strongly hydrophobic). It is similar in its composition to the C-terminal regions of the α -subunits encoded by both 'major' and 'minor' pea legumin genes ([36,37]; J. A. Gatehouse & D. Bown, unpublished work), but the actual amino acid sequences are not significantly homologous when compared by a dot-matrix homology plot (results not shown). This additional sequence is presumably responsible for the differences in physical properties between vicilin and convicillin, e.g. solubility and binding to hydroxyapatite [1]. The predicted M_r values for the mature convicillin polypeptide, and its N-terminal CNBr fragment, are not in complete agreement with those observed on SDS/polyacrylamide-gel electrophoresis. This discrepancy is a consequence of abnormal migration on electrophoresis, possibly due to the atypical amino acid composition of these polypeptides caused by the 'inserted' sequence.

Sequence

At the amino acid level, the α -subunit sequence is significantly different from vicilin, but the difference is not strong enough to be detected by a trans-

Relations

The relationship between the α -subunit sequence in pea 1 (convicillin) and the coding sequence in pea 2 (vicilin) is shown in Fig. 6. Both sequences are very similar, but the coding sequence is near the C-terminus of the vicilin sequence. However,

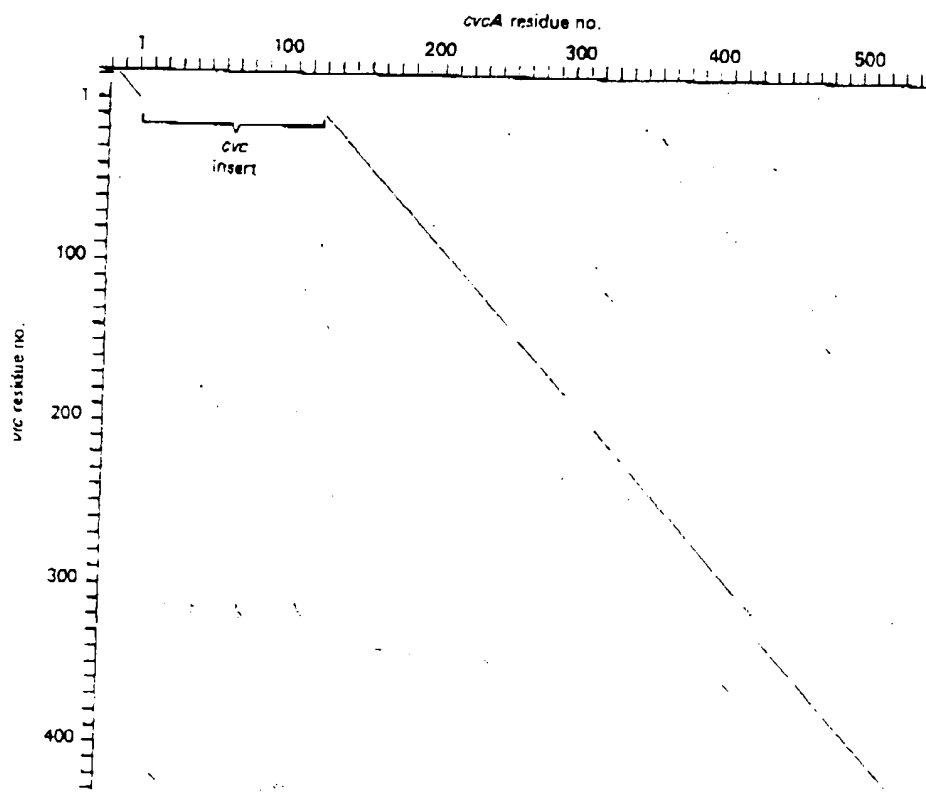


Fig. 5. Dot-matrix comparison of the amino acid sequences of vicilin (from pAD 2.1 plus *vicB*) and convicillin

Sequences were compared over a span of eight amino acids, with a minimum score of 102 using the correlation matrix given by Staden [15]

Fig. 6. Pu

The relationship between the α -subunit sequence in pea 1 (convicillin) and the coding sequence in pea 2 (vicilin) is shown in Fig. 6. Both sequences are very similar, but the coding sequence is near the C-terminus of the vicilin sequence. However,

(1), or At the nucleotide level, the inserted sequence is A + G rich, again like the C-terminal regions of legumin α -subunits; however, overall homology of nucleotide sequence in these regions is not more than marginally significant by dot-matrix comparison. No introns are present in the inserted sequence. There is no evidence of inverted repeats at the ends of the inserted sequence, nor strong evidence for direct repeats in or near the sequence itself (results not shown). The origin of this sequence is therefore unclear; it may represent a sequence inserted by a transposable element or by some other mechanism.

Relationship to vicilin-family genes in other species

The relationships of the coding sequences of vicilins in *pca*, *Phaseolus vulgaris* (phaseolin) and soya bean (conglycinin) have been extensively analysed, and part of the coding sequence of convicilin has been shown to be homologous with those of phaseolin and conglycinin [38]. Both convicilin and conglycinin have large inserted coding sequences (121 and 174 amino acids respectively) near the N-terminus of the mature protein, relative to the vicilin/phaseolin type. The inserted sequences in convicilin and conglycinin also show similarity at the nucleotide level in that both sequences are A + G-rich. However, the inserted sequences in the two genes are not

significantly homologous at either the amino acid or the nucleotide sequence level. Further, the remaining coding sequences of the two genes, although homologous, are less homologous with each other than convicilin in *pca* is with pea vicilin, suggesting that the divergence of the *pca* gene sub-families took place after the separation of pea and soya bean as species. If this is the case, the insertion events were independent of each other. Further analysis of other storage-protein gene sequences (results not shown) suggests that the insertion of hydrophilic, predominantly acidic, amino acid sequence regions is a frequent mechanism of storage protein mutation in legumes.

The flanking sequences

3' Flanking sequence. The 3' flanking sequence of *cvcA* does not show any unusual features when compared with other plant storage-protein genes.

5' Flanking sequence. Features of potential interest in the 5' flanking sequence of *cvcA* were shown by dot-matrix sequence comparisons between this gene and other plant storage-protein genes. Comparisons of the 5' flanking sequence of *cvcA* with those of conglycinin and phaseolin genes show three areas of sequence con-

'Vicilin-box' region

```

Pvu phas b      :v(106)
CC:GCCACCTCAATTTC-TTCACTTCAACACACGTCAACCTGAT:AT
Gma cgly a'     :v(88)
CC:GCCACCTCATTTTGTGTTTATTTCAACACCCGTCAAC:GCAT:CC
Psa cvcA        :v(99)
TT:GCCACCTCTATTTTGTTCATTTCAACACTCGTCAAGTTACAT:GA
:***** ** * ***** * **
: ^
: (distance to 'TATA' box)

```

Upstream region 1

```

Pvu phas b      :v(180)
GGC:TCACCATCTCAACCC:ACAC
Gma cgly a'     :v(153)
CAT:TCAC-CAACTCAACCC:ATCA
Psa cvcA        :v(152)
TAA:TCAA-CAACTCAACCC:CCGA
:*** ** *****
: ^
: (distance to 'TATA' box)

```

Upstream region 2

```

Pvu phas b      :v(257)
GGC:TGAICAGATCGCCGCTCCA:TGATG
Gma cgly a'     :v(251)
AGC:TGATCAGGATCGCCGCTCAA:GAAGAA
Psa cvcA        :v(255)
TCA:TGGTCATGATCGCCGCTCCA:TGTAA
:*** ** ***** **
: ^
: (distance to 'TATA' box)

```

Fig. 6. Putative enhancer sequences in the 5' flanking regions of *cvcA*

The three corresponding regions of high sequence homology between pea convicilin (*Psa cvcA*), *Phaseolus vulgaris* phaseolin b (*Pvu phas b*) and soya-bean conglycinin a' (*Gma cgly a'*) gene 5' flanking sequences are given. Bases the same in all three sequences are indicated by an asterisk. Homologous regions around the transcription start and the 'TATA' box are not shown.

ervation besides the 'TATA' box promoter element (considered previously): the conserved regions are shown in Fig. 6. There is also a conserved region around the transcription start, which has an obvious functional role, and a possible further conserved region of approx. 15 bases, at 30–50 bases 5' to the 'TATA' box. This latter region is not as well conserved or defined as other regions, but does include the putative CCAAT sequences of phaseolin and conglycinin [39].

The 'vicilin box' region [39] in all three genes is in a similar position (approx. 100 bases 5' to the 'TATA' box), and is strongly homologous; it can be divided into two regions, separated by 11–12 bases of T-rich sequence. The 5' region is a highly conserved C-rich sequence (GCCACCTC), whereas the 3' region is more typical of the 5' flanking sequence as a whole (TTCAACACNCGTCAANNNTG/ACAT). It has been suggested that this region, present also in pea vicilin genes, is involved in determining tissue-specificity of expression of the gene family [39]. The other two conserved regions are approx. 150–200 bases and 250 bases 5' to the 'TATA' box, like the 'vicilin box', both seem to have a highly conserved C-rich core sequence (CTCAACCC and GATCGCCGC respectively) and are associated with less highly conserved sequence more typical of the 5' flanking sequence as a whole. The hypothesis that such C-rich sequences are acting as 'enhancers' of gene expression may be advanced, and is supported by the observation that the 'vicilin-box' C-rich sequence is present in the pea legumin gene *legA* also, and has been previously observed to be homologous with a viral enhancer sequence [39,40]. However, functional assays such as those carried out with the conglycinin gene in transgenic petunia plants [41] are needed to test this conclusion.

We thank Dr. H. Hirano, National Institute of Agricultural Resources, Tsukuba, Japan, for carrying out the automated protein sequencing, John Gilroy for performing manual protein sequencing, and Paul Preston for skilled technical assistance in DNA sequencing. We also thank Professor D. Boulter for providing departmental facilities. Financial support from the Agriculture and Food Research Council and the Science and Engineering Research Council is gratefully acknowledged.

REFERENCES

1. Croy, R. R. D., Gatehouse, J. A., Tyler, M. & Boulter, D. (1980) *Biochem. J.* **191**, 509–516.
2. Croy, R. R. D. & Gatehouse, J. A. (1985) in *Plant Genetic Engineering* (Dodds, J. H., ed.), pp. 143–268. Cambridge University Press, Cambridge.
3. Thomson, J. A. & Schroeder, H. E. (1978) *Aust. J. Plant Physiol.* **5**, 281–294.
4. Domoney, C. & Casey, R. (1983) *Planta* **159**, 446–453.
5. Casey, R., Domoney, C. & Stanley, J. (1984) *Biochem. J.* **224**, 661–666.
6. Domoney, C. & Casey, R. (1985) *Nucleic Acids Res.* **13**, 687–699.
7. Matta, N. K. & Gatehouse, J. A. (1982) *Heredity* **48**, 383–392.
8. Mahmoud, S. H. & Gatehouse, J. A. (1984) *Heredity* **53**, 185–191.
9. Ellis, T. H. N., Domoney, C., Castleton, J., Cleary, W. & Davies, D. R. (1986) *Mol. Gen. Genet.* **205**, 164–169.
10. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
11. Vieira, J. & Messing, J. (1982) *Gene* **19**, 259–268.
12. Messing, J. (1983) *Methods Enzymol.* **101**, 20–78.
13. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467.
14. Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3961–3965.
15. Staden, R. (1982) *Nucleic Acids Res.* **10**, 295–306.
16. Hopp, T. R. & Wood, K. R. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 3824–3828.
17. Kuhn, S., Anitz, H. J. & Starlinger, P. (1979) *Mol. Gen. Genet.* **167**, 235–241.
18. Rigby, P. W. J., Dieckmann, M., Rhodes, D. & Berg, P. (1977) *J. Mol. Biol.* **113**, 237–251.
19. Ellis, T. H. N., Davies, D. R., Castleton, J. A. & Bedford, I. D. (1984) *Chromosoma* **91**, 74–81.
20. McMaster, G. K. & Carmichael, G. G. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 4835–4838.
21. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* **18**, 5294–5299.
22. Thomas, P. S. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 5202–5205.
23. Favaloro, I., Treisman, R. & Karnen, R. (1980) *Methods Enzymol.* **65**, 718–749.
24. Gatehouse, J. A., Evans, I. M., Bown, D., Croy, R. R. D. & Boulter, D. (1982) *Biochem. J.* **208**, 119–127.
25. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
26. Gatehouse, J. A., Lycett, G. W., Croy, R. R. D. & Boulter, D. (1982) *Biochem. J.* **207**, 629–632.
27. Slightom, J. L., Sun, S. M. & Hall, T. C. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 1897–1901.
28. Lycett, G. W., Delauney, A. J., Gatehouse, J. A., Gilroy, J., Croy, R. R. D. & Boulter, D. (1983) *Nucleic Acids Res.* **11**, 2367–2380.
29. Delauney, A. J. (1984) Ph.D. Thesis, University of Durham.
30. Lycett, G. W., Delauney, A. J. & Croy, R. R. D. (1983) *FEBS Lett.* **153**, 43–46.
31. Messing, J., Geraghty, D., Heidecker, G., Hu, N., Kridl, J. & Rubinstein, I. (1983) in *Genetic Engineering of Plants* (Kosuge, T., Meredith, C. P. & Hollander, A., eds.), pp. 211–227. Plenum Publishing Corp., New York.
32. Von Heijne, G. (1985) *J. Mol. Biol.* **184**, 99–105.
33. Croy, R. R. D., Lycett, G. W., Gatehouse, J. A., Yarwood, J. N. & Boulter, D. (1982) *Nature (London)* **285**, 76–79.
34. Tyler, M. (1981) Ph.D. Thesis, University of Durham.
35. Higgins, T. J. V. & Spencer, D. (1981) *Plant Physiol.* **67**, 205–211.
36. Lycett, G. W., Croy, R. R. D., Shirat, A. & Boulter, D. (1984) *Nucleic Acids Res.* **12**, 4493–4506.
37. Gatehouse, J. A., Bown, D., Gilroy, J., Levasseur, M., Castleton, J. & Ellis, T. H. N. (1988) *Biochem. J.* **250**, 15–24.
38. Doyle, J. J., Schuler, M. A., Godette, W. D., Zenger, V., Beachy, R. N. & Slightom, J. L. (1986) *J. Biol. Chem.* **261**, 9228–9238.
39. Gatehouse, J. A., Evans, I. M., Croy, R. R. D. & Boulter, D. (1986) *Philos. Trans. R. Soc. London B* **314**, 367–384.
40. Lycett, G. W., Croy, R. R. D., Shirat, A. H., Richards, D. M. & Boulter, D. (1985) *Nucleic Acids Res.* **13**, 6733–6743.
41. Chen, Z.-L., Schuler, M. A. & Beachy, R. N. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8560–8564.

Received 7 May 1987; 14 October 1987; accepted 21 December 1987